On Longitudinal Item Response Theory Models: A Didactic

Chun Wang
University of Washington
Steven W. Nydick
Korn Ferry

Correspondence concerning this manuscript should be addressed to Chun Wang at:

312E Miller Hall
Measurement & Statistics
College of Education
Box 353600
Seattle, WA 98195-3600
e-mail: wang4066@uw.edu
phone: 206-616-6306

Acknowledgement: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170042 (or R305D160010) awarded to the University of Washington and the Spencer/National Academy of Education Post-doc Fellowship (2014). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Citation: Wang, C., & Nydick, S. (2019). On longitudinal item response theory models: A Didactic. *Journal of Educational and Behavioral Statistics*.

On Longitudinal Item Response Theory Models: A Didactic

-

-

Abstract

Recent work on measuring growth with categorical outcome variables has combined the item response theory (IRT) measurement model with the latent growth curve (LGC) model (e.g., McArdle, 1988) and extended the assessment of growth to multidimensional IRT models (e.g., Hsieh, von Eye, & Maier, 2010; Huang, 2013) and higher-order IRT models (e.g., Huang, 2015). However, there is a lack of synthetic studies that clearly evaluate the strength and limitations of different multilevel IRT models for measuring growth. This study aims to introduce the various longitudinal IRT models, including the longitudinal unidimensional IRT model (L-UIRT), longitudinal multidimensional IRT model (L-MIRT), and longitudinal higher-order IRT model (L-HO-IRT), which cover a broad range of applications in education and social science. Following a comparison of the parameterizations, identification constraints, strengths, and weaknesses of the different models, a real data example is provided to illustrate the application of different longitudinal IRT models to model students' growth trajectories on multiple latent abilities.

Keywords: Item response theory, latent growth curve model, overall ability, domain ability

On Longitudinal Item Response Theory Models: A Didactic

Introduction

In education, one is often interested in determining student growth. These changes can sometimes be captured by latent variable models. The latent variables, such as students' abilities, are typically measured by binary (or polytomous) responses to items. Item response theory (IRT) models are useful tools to model the relationship between the categorical outcome variables and the latent continuous traits. Recent work has extended IRT models to model changes in latent traits, leading to the family of longitudinal IRT models (e.g., Andersen, 1985; Cai, 2010b; Fischer, 1973, 1976; Huang, 2013; Hsieh, von Eye, & Maier, 2010; McArdle, Grimm, Hamagami, et al., 2009; Paek, Li, & Park, 2016; von Davier, Xu, & Carstensen, 2011; Wang, Kohli, & Henn, 2016; Wilson, Zheng, & McGuire, 2012). Within this family, models differ mainly in the following aspects: (1) the measurement model that implies the factor structure of the primary latent traits measured repeatedly, which could either be unidimensional, multidimensional (Hsieh, et al., 2010), or hierarchical (Huang, 2013); (2) the relationship of the latent traits over time, which could either be captured by a completely unstructured covariance matrix (Andrade & Tavares, 2005; Cai, 2010b; Paek, et al., 2016), or linear/nonlinear change patterns via the latent growth curve models (LGC; Bollen & Curran, 2006; Duncan, Duncan, & Strycker, 2006); and (3) whether nuisance factors are in place to account for the dependency of the same items administered over time (e.g., two-tier model, Cai, 2010b; Paek, 2016; Wang, et al., 2016).

Due to the well-known connection between IRT and categorical factor analysis (e.g., Takane & de Leeuw, 1987), longitudinal IRT models can also be discussed in structural equation modeling terms. However, IRT offers two conceptual advantages: (1) assuming item (or anchor item) parameters are the same over time to ensure longitudinal invariance of the lowest-order traits; and (2) incorporating guessing parameters into the functional form of the model.

Different forms of longitudinal IRT models were proposed by different groups of researchers, and they have all been individually demonstrated to work well; however, few studies have explored the connections among the models nor the strengths and limitations of each of them. Our goal here is to capitalize on the shared features and distinctions among various longitudinal

IRT models to provide practitioners with coherent guidelines about the conditions under which each model could be applied and/or should be preferred.

Three specific types of models will be the focus of discussion. In order of complexity, these models include the longitudinal unidimensional IRT model (L-UIRT; Wang, et al., 2016; Wilson, et al., 2012), longitudinal multidimensional IRT model (L-MIRT; Hsieh, et al., 2010), and longitudinal higher-order IRT model (L-HO-IRT; Huang, 2013). All of these models are variations of the general LGC model and the respective measurement model: the undimensional IRT model assumes that a single latent trait is measured by all the items; multidimensional IRT models posit that item responses are probabilistically determined by multiple, usually correlated, latent traits; the higher-order IRT models (de la Torre & Song, 2009; Sheng & Wikle, 2008) capture the hierarchical nature of factor structure (e.g., Huang & Wang, 2014; Sawaki, Stricker, & Oranje, 2009), whereby a general factor (such as math aptitude) informs domain-specific factors (such as algebra, geometry, calculus, or subsets thereof). These three models were selected to cover a majority of practical applications. Moreover, LGC models were chosen over an unstructured covariance matrix because LGC results in both group level and individual-level growth trajectories, which are often useful for interpreting data patterns. On the other hand, LGC introduces additional latent variables (i.e., individual intercepts and slopes) that complicates model identification constraints and requires additional guidelines for model estimation. Note that the longitudinal MIRT model with unstructured covaraince matrix of θ over time is discussed in detail in Paek et al. (2016).

In the remaining sections, we introduce the three models and explain when each model could be applied. For each model, we describe identification constraints, which can be different depending on whether some items have pre-calibrated parameters. After determining the identification requirements, we are then ready to estimate the models. Estimation presents various challenges, and we describe the available estimation methods, complications due to high-dimensionality, and possible solutions. We finally illustrate the models with a real data example.

Longitudinal IRT Models

Longitudinal UIRT Model

If only one primary latent trait is measured over time, then the simplest model, the L-UIRT model, can be applied. Let $\boldsymbol{\theta}_i$ denote the T-by-1 vector of the unidimensional trait for person i across T time points. Assume there are p fixed (denoted as $\boldsymbol{\beta}$) and q random (denoted as $\boldsymbol{\nu}_i$) effects explaining the growth pattern of θ. Then the latent growth curve model on $\boldsymbol{\theta}_i$ can be written in a general form as follows

$$\boldsymbol{\theta}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\nu}_i + \boldsymbol{\delta}_i. \tag{1}$$

In Equation (1), $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the T-by-p and T-by-q design matrices for the fixed effects and random effects, respectively. In a simple LGC model with only random intercepts and random slopes, $p = q = 2$ and $\boldsymbol{X} = \boldsymbol{Z} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{pmatrix}$. $\boldsymbol{\delta}_i$ is a T-by-1 vector of residuals. The random effects are often assumed to follow a multivariate normal distribution with a mean of 0's, and a covariance matrix of $\boldsymbol{\Sigma}_\nu$. Note the number of measurement occasions, T, can be different for each person in the LGC model, allowing for missing data by design. For simplicity, we keep T the same across persons in this paper.

For a simple linear growth model with a a single person-specific intercept and slope, we can rewrite Equation (1) as

$$\theta_i^t = \pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^t, \tag{2}$$

where π_{0i} and π_{1i} are the individual intercept and slope parameters. The individual intercepts/slopes can be further written as deviations from an overall intercept (β_0) and slope (β_1) as $\pi_{0i} = \beta_0 + \nu_{0i}$ and $\pi_{1i} = \beta_1 + \nu_{1i}$.

The latent variable described by Equation (2), θ_i^t, can be measured by responses to assessment items. Assuming that responses are binary, one can model the probability of correctly responding to any item given a particular value of the latent variable with the two-parameter logistic model (2PL). The 2PL defines the probability of examinee i correctly responding to item j by the following item response function (IRF):

$$P_j(\theta_i^t) = \Pr(Y_{ij}^t = 1 | \theta_i^t, a_j^t, b_j^t) = \frac{1}{1 + \exp[-a_j^t(\theta_i^t - b_j^t)]}, \tag{3}$$

where a_j^t and b_j^t refer to discrimination and difficulty parameters for item j administered at time t. This notation is flexible enough to accommodate item sets varying across time.

Figure 1 shows an illustrative path diagram of the longitudinal UIRT model with three hypothetical time points and three items per time point.

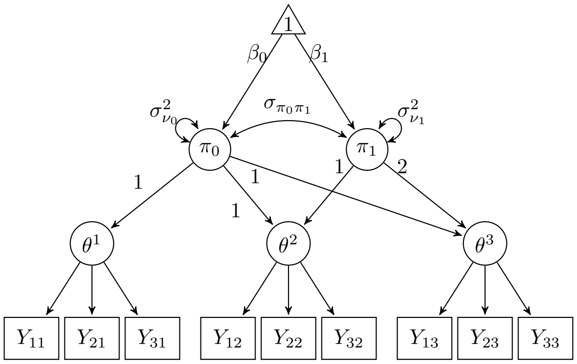

Figure 1. A path diagram for the longitudinal unidimensional IRT model with three items per time point, and three time points. π_0 represents the random intercept parameter per person, whereas π_1 represents the random linear slope parameter per person. β_0 and β_1 are the population means of π_0 and π_1, respectively.

Many large-scale educational surveys have primary measurements that differ from one occasion to another (Edwards & Wirth, 2009; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). Yet, to establish a common scale, one must either have a common set of anchor items that is shared across time or sets of anchor items that already have parameters pre-calibrated and put on a common scale (e.g., Wang, et al., 2016). Kolen and Brennan (2004) recommended that assessments should have at least 20% of items to anchor the parameters to the common scale. If enough items are linked across time, and assuming no item parameter drift, then assessments with unknown item parameters require some model identifiability constraints to be imposed.

Constraints are required to fix the mean and variance of the latent variable (ξ) at one time point (commonly $t = 1$). Given this constraint, the scale of ξ at the remaining time points will then be determined through the linking items. These constraints include

1. All of the residuals having mean 0 (i.e., $E(\delta_i^t) = 0$ for all $t = 1, \ldots, T$). This is a typical assumption in parametric regression analysis.

2. The mean of the person-specific intercept parameter being set to 0 (i.e., $\mu_{\pi_{0i}} = \beta_0 = 0$). The purpose of this assumption is to fix the mean of θ at $t = 1$ to 0.

3. The residual variance at the first time point being fixed to be a constant (i.e., $\sigma^2_{\delta_i^{(1)}} = c_1$, where c_1 is some specified constraint). This constraint indirectly fixes the variance of θ at $t = 1$.

Note that after imposing a growth curve structure on θ, θ becomes an endogenous variable in Equations (1) and (2). Hence, instead of directly fixing the mean and variance of θ (as is often desired), most SEM software packages (such as M*plus*) only allow fixing its intercept and the residual variance. The value of c_1 is arbitrary and results in the variance of θ at $t = 1$ becoming the sum of the intercept variance (i.e., $[\mathbf{\Sigma}_u]_{(1,1)}$) and c_1. When anchor items are pre-calibrated with known parameters, then only the first constraint is necessary to identify the model.

Longitudinal MIRT Model

As a multivariate extension of the L-UIRT model, the L-MIRT model combines the MIRT model with the associative latent growth curve model. The earliest version of the L-MIRT model was proposed by McArdle (1988) and called the “curve of factors model” or “CUFFS”. The CUFFS model was developed for multiple, correlated latent traits being tracked over time. For instance, the National Educational Longitudinal Study (NELS: 88) tracked students’ academic performance across three measurement occasions on four correlated cognitive scales: mathematics, reading, science, and social studies. In this case, the L-MIRT instead of L-UIRT can better recover the group-level and individual-level growth trajectories by considering all related information. Please note that name “L-MIRT” instead of “CUFFS” is used throughout the didactic for consistency with the other models’ names.

Let $\boldsymbol{\theta}_i = (\theta^1_{i1}, \ldots, \theta^1_{iK}, \ldots, \theta^T_{i1}, \ldots, \theta^T_{iK})'$ be a $KT \times 1$ vector where T denotes the number of time points and K denotes the number of correlated latent traits (i.e., dimensions) measured at each time point. Assume again that there are p fixed and q random effects per dimension. Then the general multivariate latent growth curve model can be written as

$$\boldsymbol{\theta}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\nu}_i + \boldsymbol{\delta}_i. \tag{4}$$

Similar to the notations in Equation (1), $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the $KT \times Kp$ and $KT \times Kq$ design matrices. The fixed effect, $\boldsymbol{\beta}$, is a $Kp \times 1$ vector, which is arranged in the following order: (1) the K intercepts; (2) the K slopes for the first fixed covariate; (3) the K slopes for the second fixed covariate, etc., until (p) the K slopes for the $(p-1)$th fixed covariate. This can be written in an equation as $\boldsymbol{\beta} = (\beta_{01}, \beta_{02}, \ldots, \beta_{0K}, \beta_{11}, \beta_{12}, \ldots, \beta_{1k}, \ldots, \beta_{(p-1)1}, \ldots, \beta_{(p-1)K})'$

Similarly, $\boldsymbol{\nu}_i$ is a $Kp \times 1$ vector of random effects with a covariance matrix represented by $\boldsymbol{\Sigma}_\nu$. Often, $\boldsymbol{\Sigma}_\nu$ is assumed to be a full matrix, which allows random intercepts and slopes to be correlated within and across all domains. Finally, the residuals of $\boldsymbol{\theta}_i$ are represented by $\boldsymbol{\delta}_i = (\delta^1_{i1}, \ldots, \delta^1_{iK}, \ldots, \delta^T_{i1}, \ldots, \delta^T_{iK})'$, a $KT \times 1$ random vector. The covariance matrix of $\boldsymbol{\delta}_i$, $\boldsymbol{\Sigma}_\delta$, is often assumed to be diagonal and have the following structure

$$\begin{pmatrix} \Sigma_1 & \cdots & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \Sigma_T \end{pmatrix}_{KT \times KT} ,$$

where $\Sigma_t = \text{diag}(\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_K)$ and $\boldsymbol{\Sigma}_\delta$ has T such diagonal blocks.

To be consistent with the description of the L-UIRT model, assume that each domain-level latent trait follows a simple linear trajectory without any additional covariates, which is analogous to the assumption made in the preceeding section. Then $p = q = 2$. If $T = 4$, then both

$\boldsymbol{X}$ and $\boldsymbol{Z}$ take the form of $\begin{pmatrix} I_K & 0I_K \\ I_K & 1I_K \\ I_K & 2I_K \\ I_K & 3I_K \end{pmatrix}$, where I_K is the $K \times K$ identity matrix. If nonlinear growth trajectories are considered, such as a quadratic effect of time, then $\boldsymbol{X}$ and $\boldsymbol{Z}$ would need to be updated with additional columns to account for these effects.

We can also rewrite the model by expanding Equation (4) as follows:

$$\theta_{ik}^{t} = \pi_{i0k} + \pi_{i1k} \times (t-1) + \delta_{ik}^{t}, \tag{5}$$

where π_{i0k} and π_{i1k} denote the individual intercept and slope parameters for person i on domain k. As before, the individual intercepts/slopes can be further written as deviations from an overall intercept on domain k (β_{0k}) and slope on domain k (β_{1k}), or

$$\pi_{i0k} = \beta_{0k} + \nu_{i0k} \tag{6}$$

$$\pi_{i1k} = \beta_{1k} + \nu_{i1k}. \tag{7}$$

The L-MIRT item response function takes the form of:

$$P_j(\boldsymbol{\theta}_i^t) = \Pr(Y_{ij}^t = 1|\boldsymbol{\theta}_i^t, a_j^t, b_j^t) = \frac{1}{1+\exp[-(\boldsymbol{a}_j^t)^T \boldsymbol{\theta}_i^t + b_j^t]}, \tag{8}$$

where $\boldsymbol{a}_j^t$ is a vector of discrimination parameters for item j at time t, and "T" denotes transpose. This equation is general enough to include both within-item and between-item multidimensionality structures (Recakase, 2009). Figure 2 provides an illustrative path diagram for a L-MIRT model with three measurement occasions, two domains per measurement occasions, and three items per domain. This path diagram only illustrates between-item multidimensionality.

As in the L-UIRT model, items can differ across time, as reflected by the superscript "t" on item parameters in Equation (8), but anchor items must still be embedded in the item parameter sets to link the scale. Because each domain has a potentially unique scale, anchor items must load on every domain so that the scale of θ_{ik} is linked across time for all $k = 1, \ldots, K$. As in the unidimensional case, if enough items are linked across time but all item parameters are unknown, then constraints are required to determine the scale of θ_{ik} for $k = 1, \ldots, K$. These constraints are similar to those for the L-UIRT model and include

1. All of the residuals having mean 0 (i.e., $E(\delta^t_{ik}) = 0$ for all $t = 1, \ldots, T$ and $k = 1, \ldots, K$).

2. The mean of the person-specific intercept parameters being set to 0 (i.e., $\mu_{\pi_{0ik}} = \beta_{0k} = 0$ for all $k = 1, \ldots, K$). The purpose of this assumption is to fix the mean of θ^t_{ik} at $t = 1$ to 0 for all $k = 1, \ldots K$.

3. The residual variances at the first time point being set to a constant (i.e., $\sigma^2_{\delta^1_{ik}} = c_{1k}$, $k = 1, ..., K$). As in the unidimensional case, fixing the variance of θ^t_{ik} at $t = 1$ (for all k) fixes the variances of θ^t_{ik} for the remaining time via the linking items. Moreover, θ^t_{ik} is endogenous to the model, so that the variance of θ^t_{ik} can only be constrained via its residual variance after partialing out the exogenous fixed and random effects.

As before, when anchor items are pre-calibrated with known parameters, only the first constraint must be specified to identify the model.

Longitudinal HO-IRT Model

Hierarchical factor structures often emerge in the social sciences to represent a latent construct of interest, such as intelligence (Golay & Lecerf, 2011), cognitive ability (Murray & Johnson, 2013), or personality (DeYoung, 2006). General factors are often comprised of several highly related specific factors (a.k.a., first-order factors), each of which is measured by multiple indicators (usually referred to as items). For example, in many educational assessments one is often required to report both overall proficiency for accountability purposes as well as domain-specific proficiency for diagnostic purposes. To this end, the HO-IRT model was developed by introducing a higher-order ability (de la Torre & Song, 2009; de la Torre & Hong, 2010) that relates to each of the first-order abilities. The HO-IRT model contains two levels: (1) a link between a single overall latent trait and one of several domain latent traits; and (2) a probabilistic relationship between each domain latent trait and items designed to measure that

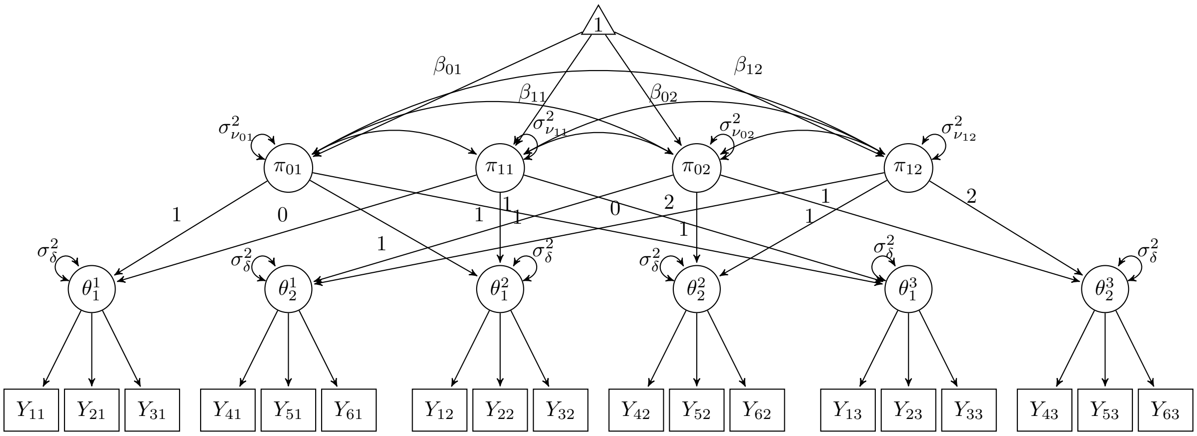

Figure 2. A path diagram for the longitudinal multidimensional IRT model with three items per domain-level trait, two domain-level traits per time point, and three time points. π_{01} and π_{02} represent the random intercept parameters per person for both domains, whereas π_{11} and π_{12} represent the random linear slope parameters per person. β_{01}, β_{12}, β_{11} and β_{12} are the population means of π_{01}, π_{02}, π_{11}, and π_{12} respectively.

domain. Specifically, let θ represent the domain latent trait underlying responses to test items, and denote ξ as the higher-order trait. Then one can hypothesize that

$$\theta_{ik} = \lambda_k \xi_i + \epsilon_{ik}, \tag{9}$$

where ξ_i is the overall ability of examinee i, θ_{ik} represents domain-specific ability $k \in \{1, \ldots, K\}$ for examinee i , λ_k indicates the relationship between domain-specific ability k and overall ability, and ϵ_{ik} is a disturbance term that can be interpreted as the domain-specific component of the ability not explainable by ξ_i. According to de la Torre and Song (2009), the residuals in Equation (9) are usually assumed uncorrelated across domains, which results in $\boldsymbol{\epsilon}_i$ (containing all of the ϵ_{ik}s) having a diagonal covariance matrix. Note that the variance of ϵ_{ik} is the unique variance of the first-order factor that is not shared by the common second-order factor. At a lower level, the probability of examinee i correctly responding to item j on domain k is defined by the same item response function (IRF) in Equation (3) except replacing ξ_i with θ_{ik}. As a result, The IRF in Equation (3) implies between-item multidimensionality which is often assumed in the HO-IRT models (e.g., de la Torre & Song, 2009; Wang, 2014). Other measurement models could also be considered based on the properties of the test.

To extend the HO-IRT model across T time points, assume the second-order factor (i.e., overall ability) follows the LGC model, as in Equation (1). Then the domain-specific ability for person i at time t would also be predicted to systematically change over time (Huang, 2013, 2015) as follows:

$$\boldsymbol{\theta}_i = \boldsymbol{\lambda}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\nu}_i + \boldsymbol{\delta}_i) + \boldsymbol{\epsilon}_i. \tag{10}$$

Equation (10) can be further understood by expanding it using a scalar equation. That is, given Equations (6) and (7), a domain-specific ability for person i at time-point t, θ^t_{ik}, would also follow a linear change over time,

$$\begin{aligned}\theta^t_{ik} = \lambda_k \xi^t_i + \epsilon^t_{ik} &= \lambda_k(\pi_{0i} + \pi_{1i} \times (t-1) + \delta^t_i) + \epsilon^t_{ik} \\ &= \lambda_k \pi_{0i} + \lambda_k \pi_{1i} \times (t-1) + (\lambda_k \delta^t_i + \epsilon^t_{ik}) \\ &= \zeta_{0ki} + \zeta_{1ki} \times (t-1) + \upsilon^t_{ik}. \end{aligned} \tag{11}$$

Notably, Equation (11) implies that the loading of the domain-specific factors on the overall factor remains the same over time, as indicated by the lack of a superscript t on λ_k. By assuming invariance of the factor structure, Equation (11) ensures that the lower-order factors carry the same meaning over time, which fulfills the "longitudinal measurement invariance" property (Chen, Sousa, & West, 2005; Liu, Millsap, West, et al., 2016). Figure 3 provides an illustrative path diagram of the L-HO-IRT model, assuming three time points, two domain-specific abilities per time point, and three items measuring each domain-specific ability.

As shown in Equations (5) and (11), the L-HO-IRT model is nested within the L-MIRT model. This is because the L-MIRT model allows for separate, potentially unrelated, individual intercept and slope parameters across each dimension (i.e., π_{i0k} and π_{i1k}). Conversely, the L-HO-IRT model restricts the domain-level intercept and slope parameters to take the predetermined structure of $\lambda_k \pi_{i0}$ and $\lambda_k \pi_{i1}$ due to the functional form of the model.

Assuming either the same sets of items are repeatedly administered or that the test includes shared items between adjacent time points for all domains, the minimum model identifiability

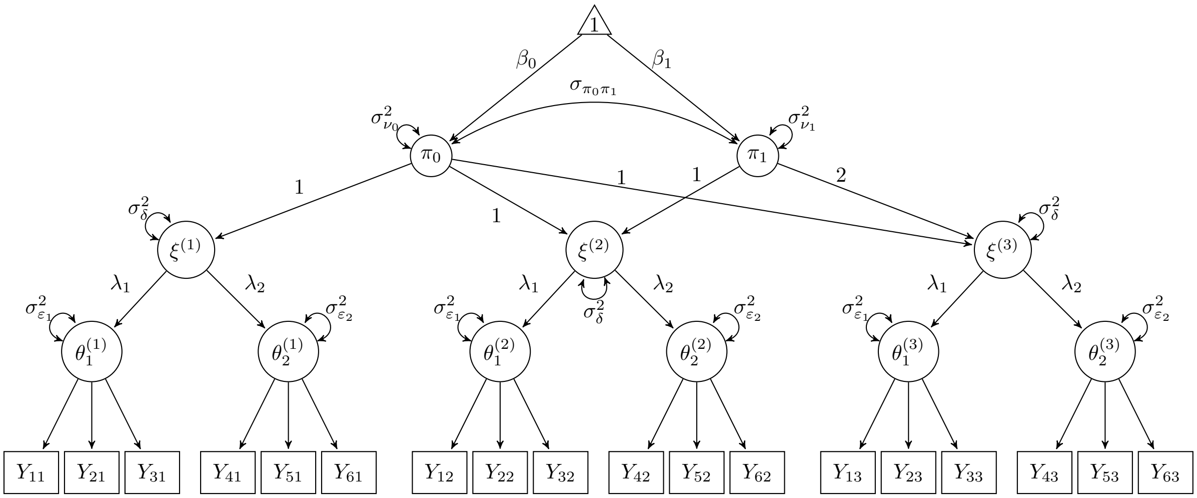

Figure 3. A path diagram for the longitudinal, higher-order IRT model with three items per domain-level trait, two domain-level traits, and three time points. π_0 and π_1 represent the random intercept and random linear slope parameters per person. β_0 and β_1 are the population means of π_0 and π_1, respectively.

constraints include:

1. All of the residuals having mean 0 (i.e., $E(\delta_i^t) = 0$ for all $t = 1, \ldots, T$).

2. The mean of the person-specific intercept parameters being set to 0 (i.e., $\mu_{\pi_{i0}} = \beta_0$). The purpose of this constraint is to specify the location of ξ^t at $t = 1$.

3. All of the residuals in the measurement model having mean 0 (i.e., $E(\epsilon_{ik}^t) = 0$ for all $k = 1, \ldots, K$ and $t = 1, \ldots, T$) in Equation (9). This assumption is typical for a factor regression model and made in de la Torre & Song (2009).

4. The residual variances at the first time point being set to a constant. (i.e., $\sigma^2_{\epsilon^1_{ik}} = c_k^1$ where c_k^1 is a user-specified constant). This constraint is necessary to establish the scale of the θ's in the model. Justification for this constraint is similar to justification for the similar constraint in the L-UIRT and L-MIRT models and is due to $\boldsymbol{\theta}_i^t$ being endogenous to the model, so that its variance can only be fixed indirectly by setting its residual variance. Only the variance at a single time point needs to be fixed, as the variance of $\boldsymbol{\theta}_i^t$ at the remaining

time points are determined via the linking items.

5. One of the loading parameters, λ_k for some some k $(k = 1, \ldots, K)$ being set to a constant, assuming that λ_k is invariant over time. The remaining $(K - 1)$ loading parameters are freely estimable.

The first two constraints are essentially the same as the first two constraints for both the L-UIRT model and the L-MIRT model, described earlier. The remaining constraints are unique to the L-HO-IRT model. The last constraint is similar to the "reference indicator" constraint in factor analysis. That is, the variance of a factor can be determined by fixing the loading of one marker indicator. Here, the "marker indicator" is one of the first-order factors, θ_{ik} for some k $(k = 1, \ldots, K)$, and the "factor" is ξ_i. Readers of de la Torre and Hong (2010) may notice that they imposed a different constraint for the same purpose, namely

$$\text{var}(\epsilon^t_{ik}) = 1 - \lambda^2_k. \tag{12}$$

They argued that the variance of θ_{ik} is typically assumed to be 1, and the assumption from Equation (12) results in a variance of ξ_i also assumed to be 1. Thus, by way of this constraint, both the first-order and second-order factors would be on the same scale. The motivation of de la Torre and Hong (2010) are not relevant to our current discussion, as the variance of θ_{ik} is not assumed to be a constant over time (and might have good reason not to be given the type of change observed). If requiring standardized loading parameters, one could calculate a simple linear transformation of λ_k, that is $\lambda^\star_k = \lambda_k \times \frac{\sigma_{\xi^t}}{\sigma_{\theta^t_k}}$. Moreover, in M*plus*, the equality constraints in (12) can only be specified with maximum likelihood estimation (MLE) but not with the Bayesian estimation option. Note that when anchor items are pre-calibrated with known parameters, then only the first, third, and last constraints are necessary to identify the model.

Applications of the Models

Applying one of the above models versus another depends mostly on the hypothesized factor structure of the latent traits. Higher-order models are often applicable in contexts where a measurement instrument assesses several related constructs that can be accounted for by one or more underlying second-order factors (Chen, et al., 2006). For instance, a common scale to measure "quality of life" is composed of four *subscales* that each presume to measure a distinct first-order factor: mental health, cognition, vitality, and health worry (Chen et al., 2006). The covariance between each pair of first-order factors can be explained by a higher-order factor, which is usually called "global quality of life." Similarly, educational measures are often constructed to assess several, separate but correlated, content domains that can be partially explained by a more general ability. For instance, a mathematics test may have items measuring numerical computation skills and data analysis skills (Reckase, 2009, p. 232). Both of these are examples of content-based multidimensionality rather than strict construct-based multidimensionality.

In practice, one cannot typically distinguish between content multidimensionality and construct multidimensionality because content-based subscales often measure distinct constructs. Yet certain content-based domains sometimes have exceedingly high correlations, implying that these domains essentially measure the same skill or construct (Reckase, 2009). In cases like these, one should always provide evidence that combining domains makes substantive sense or yields a better fit than keeping those domains separate.

Although a correlated-factor MIRT model will always fit data generated from the higher-order IRT model, the higher-order model has at least four advantages for being preferred in practice: as compared with the correlated-factor MIRT model, the HO-IRT model (1) parsimoniously explains the covariance between lower-order factors (Gustafsson & Balke, 1993; Rindskopf & Rose, 1988); (2) separates the variance in the lower-order factors shared by the common higher order factor from the unique variance of the lower-order factors; (3) simplifies model estimation due to the exploitation of the dimension reduction technique (as described in the next section); and (4) allows for potential construct shifts over time.

To elaborate on the last point, assume teachers want to track students' ability in a general subject area, such as math knowledge. If math knowledge is a unidimensional trait, it can be measured directly by a set of items, and if the teacher is not interested in measuring any specific sub-areas of mathematics, then the L-UIRT model is sufficient. However, math knowledge might relate to a number of specific content areas that teachers might also wish to track. For example, Table 1 presents the content coverage of the mathematics common core domains across five domains. The domains (such as Domain 5 “Geometry” and Domain 4 “Measurement and Data”) are expected to be taught and developed in every grade from Kindergarten-4. Student growth in these domains can be tracked across all five grades. However, the required content coverage shifts from grade to grade, and many domains only appear in limited grades. For instance, Domain 1 (“Counting and Cardinality”) is expected to be assessed only in Kindergarden, whereas Domain 6 (“Numbers and Operations-Fractions”) does not emerge until Grade 3. In these cases, the L-MIRT model and L-UIRT model overlook crucial details. In particular, the longitudinal MIRT model (Hsieh, et al., 2010) essentially assumes a constant set of traits measured over time. For this relatively straightforward example, the domains are designed to change over time.

However, when indeed the same of domains are measured overtime, the L-MIRT model is preferred because the L-HO-IRT model is parametrically more restricted than the L-MIRT model. That is, any growth patterns in the lower-level traits that can be captured with the L-HO-IRT model can ultimately be captured with the L-MIRT model. Yet if the multidimensional (lower-level) constructs each change differently over time, then the L-HO-IRT model would no longer fit the data, and one should use the L-MIRT model. For instance, if certain domain-level traits grow linearly, whereas others grow in a piecewise fashion, then one should no longer use the L-HO-IRT model due to the restrictions implicit in Equation (10). On the other hand, the L-MIRT model can handle different growth patterns if needed.

When assessing change over time, one must consider whether the measures retain measurement invariance. Often, practitioners use the exact same scale on multiple occasions. This practice can ensure that identical constructs are continuously assessed and that the metric of measurement remains the same over time. However, out of necessity, scales often differ across repeated measurements due to the need for “developmentally appropriate measures” (Widaman, Ferrer, & Conger, 2010). Adjusting the scale to consider the typical range of traits over repeated measurements can help avoid ceiling and floor effects (Embretson, 2006, 2007; May & Nicewander, 1998).

Determining whether the same construct, measured by multiple indicators, has the same meaning and metric over time falls under the rubric of measurement invariance (Widaman et al., 2010), and is often referred to, especially in a longitudinal setting, as longitudinal invariance. The factorial invariance of longitudinal measures is paramount in evaluating the change in behavior over time (Meredith & Tisak, 1990; McArdle, 2001; McArdle & Hamagami, 2001; Widaman & Reise, 1997). Using the same set of items, or a set of anchor items (Grimm et al., 2013) partially satisfies longitudinal invariance. A thorough examination of longitudinal invariance is beyond the scope of this article. Interested readers can refer to Teresi (2006), Isiordia and Ferrer (2018), Liu, Millsap, West, et al. (2016) for details regarding invariance assumptions of L-UIRT, L-MIRT (i.e., CUFFS), and L-HO-IRT, respectively.

Table 1

Mathematics common core domains by grade (K-4)

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4
Domain 1	Counting and Cardinality				
Domain 2	Operations and Algebraic thinking	Operations and Algebraic thinking	Operations and Algebraic thinking	Operations and Algebraic thinking	Operations and Algebraic thinking
Domain 3	Number and Operations in Base 10	Number and Operations in Base 10	Number and Operations in Base 10	Number and Operations in Base 10	Number and Operations in Base 10
Domain 4	Measurement and Data	Measurement and Data	Measurement and Data	Measurement and Data	Measurement and Data
Domain 5	Geometry	Geometry	Geometry	Geometry	Geometry
Domain 6				Number and Operations-Fractions	Number and Operations-Fractions

Model Estimation

Within the general framework of structural equation modeling (SEM), the longitudinal IRT models can be viewed as a multilevel latent growth curve model with the lowest level represented by categorical indicators. Unsurprisingly, the longitudinal IRT models can also be motivated from the framework of generalized linear models (McCullagh & Nelder, 1989), a conceptualization favored within biostatistics. The most common methods for estimating multilevel models are based on integrating the likelihood over the distribution of random effects, which is often referred to as marginal likelihood estimation. For instance, in the L-HO-IRT model, the overall and domain-specific latent abilities as well as the latent intercepts and slopes represent the random effects over which to integrate. Because analytical integrals often do not exist for these types of models, researchers frequently adopt one of two classes of methods. One could either approximate the integrand analytically or evaluate the integral via numerical approximation. The first approach includes Laplace's method of linearizing the integrand via a sixth-order Taylor series approximation (called 'Laplace 6' in Raudenbush, Yang & Yosef, 2000) as well as quasi-likelihood methods such as Marginal Quasi-Likelihood (MQL; Goldstein, 1991; Goldstein & Rasbasch, 1996) and Penalized Quasi Likelihood (PQL; Breslow & Clayton, 1993; Laird, 1978). Because the performance of PQL and MQL depends on the validity of a normal approximation, these methods tend to perform poorly when the observed data are markedly non-normal (Rodriguez, & Goldman, 1995; Tuerlinckx, Rijmen, Verbeke, & Paul De Boeck, 2006) and are thus typically not recommended for use in IRT models with binary responses. The second approach includes ML using Gauss-Hermite quadrature, adaptive quadrature, and simulation methods (Bauer & Sterba, 2011), such as the Monte Carlo EM algorithm (Wang & Xu, 2015).

However, ML estimation via the EM-algorithm is known to converge slowly in many applications (e.g., Meng & van Dyk, 1997) and is computationally intensive when the number of latent variables is large. Bayesian estimation using Markov chain Monte Carlo (MCMC) with diffuse (or noninformative) priors (Patz & Junker, 1999) is an alternative to EM (Huang, 2013; Wang & Nydick, 2015) and is usually preferred for complex models.

All of the above estimation methods are based on full-information, in that the likelihood is constructed directly from the raw response pattern. Alternatively, one could adopt limited information estimation methods, such as modified weighted least squares (WLS) estimation. Rather than basing the likelihood on the complete response pattern, modified WLS estimates model parameters via the first four moments of the response contingency table. By avoiding the time-consuming numerical integration or sampling steps of the full-information methods, WLS leads to much faster convergence. However, WLS is known to yield inaccurate estimation with small sample sizes or large amounts of missing data (e.g., Forero & Maydeu-Olivares, 2009). Moreover, the parameter estimates from WLS are not as efficient as a full-information method (Muthen & Asparouhov, 2015). Given these limitations, WLS is not discussed further in this paper.

In the following subsections, we describe estimating the longitudinal IRT models in M*plus* with ML or MCMC methods. M*plus* software was chosen due to being widely used in social science research. Other IRT estimation software packages, such as *flex*MIRT (Cai, 2017; see Paek et al. 2016, for details on how to estimate similar models to those described in this paper), or general-purpose estimation packages, such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), should also be able to recover longitudinal IRT-based model parameters. Interested readers could refer to Curtis (2010) or Isiordia and Ferrer (2018), which present BUGS code and R code (using the "lavaan" package, see Rosseel, 2012), respectively, for estimating a subset of longitudinal IRT models. Details of estimating longitudinal IRT models using WLS are explained in Wang et al. (2016).

Maximum likelihood estimation

When using M*plus*, one must specify the model estimation method in the `ANALYSIS` section of the input script. If estimating IRT-based item parameters with maximum likelihood estimation, include the following `ANALYSIS` statement:

```
ANALYSIS: TYPE = GENERAL;
          ESTIMATOR = MLR;
          LINK = LOGIT;
          INTEGRATION = MONTECARLO;
```

As indicated in the last line of the previous statement, we recommend using M*plus*'s `MONTECARLO` integration routine for the numeric integration. Without including the `INTEGRATION` line, M*plus* would default to use rectangular (trapezoid) numerical integration with either 15 adaptive quadrature points per dimension, or 30 to 50 non-adaptive quadrature points per dimension (Chapter 14, M*plus* User Guide). Although adaptive numeric integration is computationally faster, if the data have outliers or non-normally distributed latent traits, it may yield unstable results. If estimating a model with 1-3 dimensions of integration, the default quadrature-based numerical integration algorithm usually results in precise estimates. Conversely, `MONTECARLO` integration does not yield as accurate estimates of parameters for low dimensions of integration but is much more efficient for higher dimensional integration.

Table 2 illustrates the dimensions of numeric integration for each of the three models with values in parentheses assuming that $T = 4$, $K = 5$, and $q = 2$. As shown in Table 2, the number of continuous latent variable per model (the second column in Table 2) is simply the number of latent factors (including the first-order and second-order latent traits) plus the number of random effects (the person-specific intercepts and slopes). The dimensions of integration (the third column in Table 2) include only those factors that have categorical indicators (the θs) as opposed to higher-level factors (the ξs) or random effects. According to the M*plus* User Guide (p. 527), closed form solutions may exist for integrating out latent factors with continuous indicators, such as the ξs or random effects, so that the numerical integration approximation is no longer needed. Nonetheless, the number of dimensions of integration for all three longitudinal models is prohibitively large.

The right-most column in Table 2 indicates the dimensions of integration if using an analytic dimension reduction technique. Analytic dimension reduction is often used to rearrange terms in the marginal likelihood integral to yield a series of integrals, each of much lower dimension than the original integral (Gibbons & Hedeker, 1992; Rijmen, Vansteelandt, & de Boeck, 2008; Cai, Yang, & Hansen, 2011). Applying a dimension reduction technique to the L-UIRT model, rewrite Equation (3) as

$$\frac{1}{1+\exp[-a_j^t(\theta_i^t - b_j^t)]} = \frac{1}{1+\exp[-a_j^t(\pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^t - b_j^t)]}. \tag{13}$$

If assuming that δ_i^t's are uncorrelated across pairs of time points, then one need only integrate out

π_{0i}, π_{1i}, and δ_i^t, resulting in a 3-dimensional integral, for a given item (Paek, et al., 2016). The same arguments lead to a similar dimension reduction solution to the L-MIRT model. The results for the L-MIRT model in Table 2 is based on the assumption that the residual covariance matrix of $\boldsymbol{\delta}_i$ is a diagonal matrix [1].

The L-HO-IRT model has a different dimension reduction solution given the addition of the higher-level trait. First, write the HO-IRT item response function as

$$\frac{1}{1+\exp[-a^t_{j_k}(\theta^t_{ik}-b^t_{j_k})]} = \frac{1}{1+\exp\{-a^t_{j_k}[\lambda_k(\pi_{0i}+\pi_{1i}\times(t-1)+\delta^t_i)+\epsilon^t_{ik}-b^t_{j_k}]\}}, \qquad (14)$$

where a_{j_k} and b_{j_k} denote item parameters for item j measuring domain k. In Equation (14), the only additional random effect to integrate out of the likelihood equation is ϵ^t_{ik}. Because all ϵ^t_{ik}'s are assumed uncorrelated across time, then generalized dimension reduction yields a 4-dimensional integral (π_{0i}, π_{1i}, and δ^t_i as before, as well as ϵ^t_{ik}). Note that this dimension reduction technique can only be applied if the residuals from the growth curve model, δ^t_i, are uncorrelated across time. If estimating models with correlated residuals (such as an autoregressive model), this dimension reduction technique can no longer be applied.

Table 2
Number of continuous dimensions and dimensions of numerical integration for different models and methods (T denotes the number of time points, K denotes the number of lower-order latent traits, q denotes the number of random effects)

Models	num. of continuous latent variables	dimensions of numerical integration (M*plus* default)	dimensions of numerical integration (analytic dimension reduction)
L-UIRT	$T+q$ (6)	T (4)	$q+1$ (3)
L-MIRT	$T\times K+q\times K$ (30)	$T\times K$ (20)	$q\times K+1$ (11)
L-HO-IRT	$T\times K+T+q$ (26)	$T\times K$ (20)	$q+2$ (4)

Advantages of estimating parameters using the EM algorithm, as compared with Bayesian methods, in M*plus* include: (1) being able to estimate the three-parameter logistic (3PL) model rather than only being able to estimate one or two parameter normal ogive models; (2) providing

[1] If, on the other hand, the residual covariance matrix of $\boldsymbol{\delta}_i$ is a block diagonal matrix, allowing the residuals from different latent traits to correlate at a given time point, then the dimensions of numerical integration would be $(q+1)\times K$.

comparative model fit indices such as AIC and BIC; and (3) being able to impose equality constraints on model parameters. Note that these limitations of Bayesian methods are not necessarily inherent to the methods themselves, only to the application of those methods in M*plus*. Due to the high-dimensional integration, we have had more success estimating the longitudinal IRT models with the MCMC option in M*plus*. Researchers and practitioners should always keep in mind complexity and feasibility when choosing a model and corresponding estimation algorithm.

Markov chain Monte Carlo

If estimating IRT-based item parameters with MCMC, include the following `ANALYSIS` statement:

```
ANALYSIS: ESTIMATOR = BAYES;
          CHAINS = 1;
          FBITER = 50000;
          POINT = MEAN;
```

In the above statement, the `FBITER` line denotes the fixed number of iterations for each Markov chain (i.e., the chain length). If `FBITER` is not specified, the chain will stop once convergence is reached with the default convergence criterion being a potential scale reduction (PSR; Gelman & Rubin, 1992) at or below 1.05 (see M*plus* user guide, 1998-2012, p. 640). After 50,000 iterations, `POINT = MEAN` indicates that the posterior mean will be used as the point estimate of the model parameters.

The next section provides a real data example of applying M*plus* (version 8 used in this study) to estimate parameters of data that fit the longitudinal IRT model. A corresponding simulation study, demonstrating parameter recovery of the three longitudinal IRT models, are included as an online appendix to this paper.

A Real Data Example

The current section applies the three longitudinal IRT models to a real data example. The purpose of this demonstration is to illustrate the potential application of each model as well as the information each model provides to researchers and practitioners. For this purpose, we adopted and analyzed a series of math assessments that students in one Midwest state took between 2009 and 2012. These students were assessed in each of grades 3-6 using a five dimensional, simple-structure test with pre-calibrated item parameters. The five dimensions had been termed "number and operation", "geometry and spatial sense", "data analysis, statistics, probability", "measurement", and "algebra, functions, and patterns", respectively. Students took 57 items in 2009 (with 23, 9, 7, 11, and 7 items, respectively, measuring each dimension) and 52 items in each of the three subsequent years (with 23, 9, 7, 11, and 7 items, respectively, measuring each dimension). After initial data cleaning, only $N = 327$ students had a complete set of mixed responses (i.e., including both correct and incorrect responses) for sets of items on each dimension at every time point[2].

Due to different sets of items being administered in each year, common-item linking is not possible. However, pre-calibrated anchor items were embedded within each of the five dimensions across all four years and are all on the same scale. Because of fixing known anchor items, many of the identifiability constraints need not be explicitly specified (see the model description section for additional details). Only λ_1 in the L-HO-IRT model must still be specified, and we set $\lambda_1 = 1$ to fix the scale of ξ. All growth models were assumed to have only random intercepts and slopes (see the spaghetti plots below for linearity of time on θ and ξ). Moreover, all responses were assumed to conform to the 2PL IRT model.[3] For estimation, a MCMC algorithm was run in M*plus* with a Markov chain length (`FBITER`) fixed to 30,000 with the first half of the iterations discarded as burn-in by default. In all cases, the PSR for all model parameters were below 1.03, implying successful chain convergence.

To evaluate global model fit in Bayesian models with categorical outcome variables, M*plus* provides the Bayesian posterior predictive p-value (Muthén, 2010; Kaplan & Depaoli, 2012). In our case, the Bayesian p-value for the L-UIRT, L-MIRT, and L-HO-IRT [4] models were estimated to be .103, .081, and .106, respectively, implying that all three models yielded acceptable global

[2] The complete data set is available for download on www.placeholder.com

[3] M*plus* can estimate 3PL model parameters using only the MMLE/EM algorithm, which becomes exceedingly slow when the number of integration dimensions is large, such as in the L-MIRT or L-HO-IRT models considered in this paper.

[4] Originally, we ran the model allowing λ_2 to λ_K to differ across time. Relaxing the invariance assumption resulted in a posterior predictive p-value changed by .001. Because imposing an invariance assumption still yields a p-value $>$ .05, we decided to base our results and discussion on the invariance model.

fit. Note that other Bayesian software packages, such as JAGS (Plummer, 2003) provides the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linden, 2002) for model comparison. M*plus* does not yet include DIC for models with categorical indicators.

Table 3
Structural model parameter estimates for three different models

Models	NP	Fixed effects $\begin{pmatrix}\beta_0 & \beta_1\end{pmatrix}$	Random effects $\begin{pmatrix}\sigma^2_{\pi_{0i}} & \\ \sigma_{\pi_{0i}\pi_{1i}} & \sigma^2_{\pi_{1i}}\end{pmatrix}$	Others
L-UIRT	275	$\begin{pmatrix}-.653 & .472\end{pmatrix}$	$\begin{pmatrix}.081 & \\ -.008 & .005\end{pmatrix}$	$\boldsymbol{\sigma}^2_{\delta_i} = \begin{pmatrix}.048 & .063 & .046 & .015\end{pmatrix}$
L-MIRT	351	$\begin{pmatrix}-.652 & .509 \\ -.633 & .428 \\ -.421 & .356 \\ -.659 & .444 \\ -.795 & .524\end{pmatrix}$	$\begin{pmatrix}.145 & .014 \\ .255 & .039 \\ .267 & .038 \\ .169 & .042 \\ .120 & .017\end{pmatrix}$	$\boldsymbol{\sigma}^2_{\delta_i} = \begin{pmatrix}.051 & .087 & .052 & .019 \\ .063 & .066 & .025 & .037 \\ .032 & .048 & .009 & .029 \\ .035 & .055 & .042 & .031 \\ .054 & .010 & .031 & .013\end{pmatrix}$
L-HO-IRT	299	$\begin{pmatrix}-.702 & .514\end{pmatrix}$	$\begin{pmatrix}.102 & \\ -.009 & .007\end{pmatrix}$	$\boldsymbol{\lambda} = \begin{pmatrix}1^\star \\ .768 \\ .734 \\ .882 \\ .935\end{pmatrix}$ $\boldsymbol{\sigma}^2_{\delta_i} = \begin{pmatrix}.052 & .071 & .061 & .015\end{pmatrix}$ $\boldsymbol{\sigma}^2_{\epsilon_i} = \begin{pmatrix}.028 & .031 & .016 & .017 \\ .121 & .091 & .088 & .044 \\ .018 & .031 & .013 & .011 \\ .059 & .020 & .018 & .034 \\ .059 & .012 & .013 & .008\end{pmatrix}$

1. “np” denotes the number of free parameters in each model.
2. The covariances between random intercepts and random slopes from the L-MIRT model are omitted to save space because they are between -.01 and .01.
3. “*” denotes a fixed constant.

Table 3 presents the parameter estimates from the three longitudinal IRT models. Because of fixing $\lambda_1 = 1$ in the L-HO-IRT model, parameter estimates from this model may not be on the same scale as those from the L-UIRT and L-MIRT models. Even though parameter estimates are not strictly comparable across models, we can still make some general statements based on Table 3: (1) the fixed effect of time is positive, implying an increase in average ability over time; (2) the intercept and slope combined variances (i.e., $\boldsymbol{Z}\boldsymbol{\Sigma}_\nu\boldsymbol{Z}^T$, where $\boldsymbol{Z}$ is the design matrix defined in Equation (1), indicating the dependent variable variance explained by random effects) greatly exceed the residual variance in the growth part of the model (i.e., $\boldsymbol{\sigma}^2_{\delta_i}$), which evidences the linear functional form being sufficient to capture the latent growth pattern; (3) the random intercepts vary more than the random slopes, and there is a moderate negative correlation (of -.3 to -.4) between random intercepts and random slopes, implying larger differences in initial ability than in growth rates. This moderate negative correlation between initial state and growth is interesting and implies that the gap between high and low performing students decreases over time. Even though one cannot directly compare parameter estimates from the L-MIRT and L-HO-IRT models, the intercept variances being larger for domains two and three in the L-MIRT model (i.e., .255 and .267 in Table 3) is consistent with the λs being relatively lower for these two domains (i.e., .768 and .734) in the L-HO-IRT model. Thus, estimation patterns persist regardless of lack of direct comparability of parameter magnitudes.

In contrast to the L-HO-IRT model, the L-UIRT and L-MIRT models can be directly compared in this case due to anchor items setting the scale for the lower-order traits. From Table 3, one can see that averaging the intercepts and slopes from the L-MIRT model leads to estimates similar to those from the L-UIRT model. Yet the variance of the intercept and slopes from the L-MIRT model are much larger, implying that evaluating individual performance at the domain level leads to higher variability than assuming that responses are all generated from a single, common trait. That said, if a test is constructed across several domains, considering domain-level growth patterns may reveal subgroup differences otherwise diminished if assuming responses came entirely from a unidimensional trait.

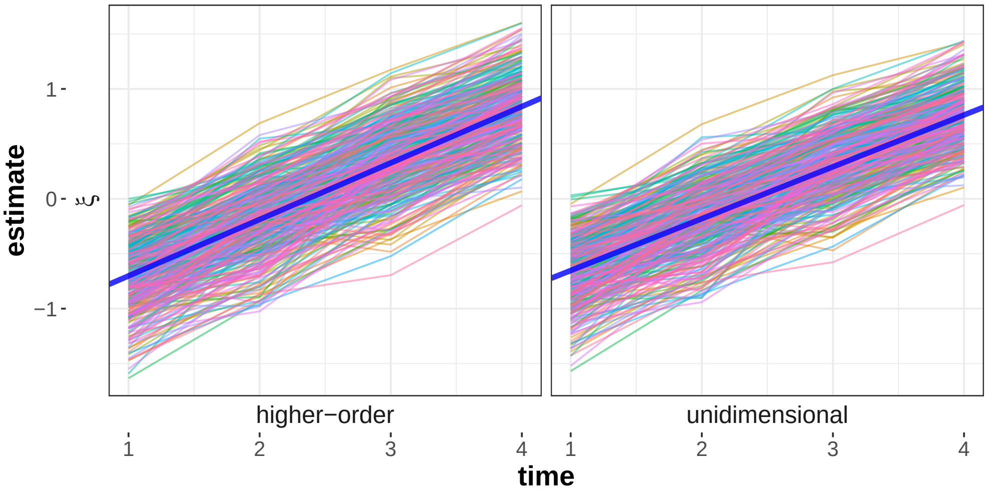

Figure 4. A spaghetti plot, illustrating the linear trend of ξ (overall-level ability) on math between grades 3–6 for $N = 327$ students. The left panel is obtained from the L-HO-IRT model, and the right panel is from the L-UIRT model. The bolded, slanted line in the center of the spaghetti depicts the estimated fixed effect of time.

Figure 4 presents a spaghetti plot of the overall ability across time for $N = 327$ students using the L-HO-IRT model (left) and the L-UIRT model (right). Unsurprisingly, the lines in the right panel are slightly closer together than the lines in the left panel, which is consistent with the results in Table 3 that the variance of the random slopes is slightly higher from the L-HO-IRT model. Figure 5 presents the spaghetti plot of the domain-specific abilities across time using the L-HO-IRT model (upper) and the L-MIRT model (lower). As shown in Figure 5, aside from minor differences, the overall growth lines and the individual growth trajectories from both models exhibit similar patterns. One anomaly worth mentioning is that the individual growth curves from the L-HO-IRT model tend to fluctuate quite a bit more than the growth curves from the L-MIRT model. The L-MIRT model growth curves (for all but $k = 1$ and $k = 4$) tend to follow strict lines. This result is due to where the growth trajectory is imposed. With respect to the L-HO-IRT model, the growth trajectory is fit to the θ's only indirectly (due to the θ's relationship with ξ) as reflected in Equation (11). Because of this indirect effect, the residual variance of θ ($\sigma^2_{\upsilon^t_{ik}} = \lambda^2_k \sigma^2_{\delta^t_i} + 1 - \sigma^2_{\epsilon^t_{ik}}$) could be large, and the individual growth trajectories might exhibit some departure from a strict line. Conversely, with respect to the L-MIRT model, a growth line is imposed directly on the individual θ's (see Equation 5). Due to a small estimated residual variance in the θ's (between 0.007 and 0.088), the domain abilities were estimated to be close to the trajectory line. Note that the figures reinforce linearity in the average growth pattern over time, which was implied earlier by comparing the slope/intercept variances to the residual variances from Table 3.

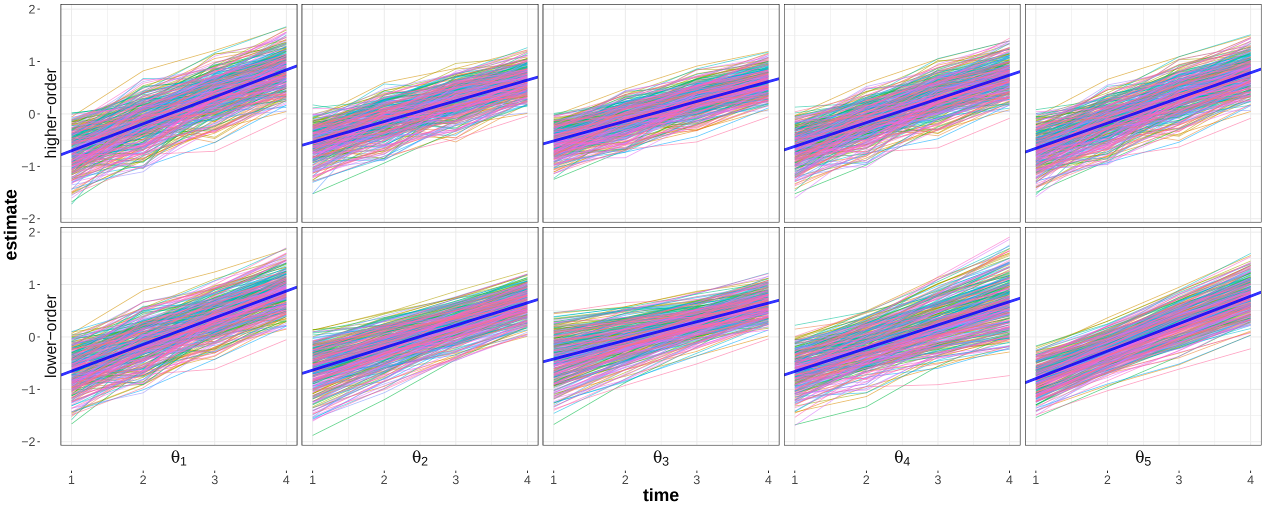

Figure 5. A spaghetti plot, illustrating the linear trend of θ (domain-level ability) on math between grades 3–6 for $N = 327$ students. The upper panels are obtained from the L-HO-IRT model, and the lower panels are from the L-MIRT model. The bolded, slanted line in the center of the spaghetti depicts the estimated fixed effect of time.

Conclusion

Many teachers, administrators, and policy makers require the measurement of student growth. Teachers can use estimated growth to modify lesson plans based on strategies of improvements. Administrators can use estimated growth to examine school performance and help make budgetary decisions. In either case, one must ensure estimates are accurate across several, possibly correlated, ability dimensions. Several longitudinal IRT models haven been proposed for different purposes. These longitudinal IRT models all share the same form and contain two components: (1) an IRT measurement model for each measurement occasion; and (2) a latent growth curve (LGC) model imposed on the latent trait, quantifying the intraindividual developmental trajectories. In this paper, we reviewed three specific types of longitudinal IRT models with the goal of demonstrating appropriate applications of these models for longitudinal assessment. We also illustrated fitting different models with a commonly used software package.

Among the three models, the L-UIRT model is the simplest and has been the most extensively studied in the literature (e.g., Andersen, 1985; Embreston, 1991, Grimm, et al., 2013; McArdle, et al., 2009; von Davier, et al., 2011; Wang et al., 2016 Wilson, et al., 2012). In contrast to the L-UIRT model, which tracks change in a unidimensional latent trait, the L-MIRT model describes change in multiple, correlated latent traits (see Paek et al., 2016). Compared to models that directly model change in the lower-level abilities, the L-HO-IRT model includes two unique features. First, because the HO-IRT model captures the hierarchical nature of learning, the L-HO-IRT model simultaneously models the growth trajectories of both overall and domain-specific abilities. Second, as described earlier in the paper, the L-HO-IRT model allows for a shift in domain coverage over time, as long as one carefully verifies the second-order longitudinal invariance requirement (e.g., Chen, Sousa, & West, 2005; Liu, et al., 2016). Allowing for a shift in the domain coverage over time is extremely important in educational measures, as one typically finds more advanced domains added and basic domains eliminated as students complete more schooling. Furthermore, a higher-order model allows one to find trends at the individual, domain level. Domain level information can hint at particular academic subjects that improve the most over particular grades. For instance, in our real data example, θ_1 and θ_5 tended to improve the most over time, and θ_3 tended to improve the least (assuming, of course, that the location and scale across dimensions are comparable). With a longitudinal HO-IRT model, one can obtain estimates of overall trends as well as delve into individual dimensions underlying complex assessments.

In terms of model estimation, we provided a thorough discussion of the analytical dimension reduction techniques that are available to alleviate high-dimensional integration challenges of marginal maximum likelihood estimation. Even after dimension reduction, the number of integration dimensions can still be high. In this case, the Metropolis-Hastings Robbins Monro algorithm (Cai, 2008, 2010) or the MCMC algorithm can be used in lieu of MMLE via EM. Given that the L-MIRT and L-HO-IRT are less studied in the literature, a simulation study was conducted to provide a thorough quality-control check on the precision in estimating model parameters (refer to the online supplementary file for details of the simulation, which evaluated the recovery of both structural parameters and individual latent traits/growth parameters). When examining simulation results, all model parameters were adequately recovered, and the generating model evidenced adequate model fit. Even with the supporting evidence from the simulation study, interested users of the L-HO-IRT and L-MIRT models should keep in mind that both of these models should only be applied when there are sufficient items per domain, otherwise the domain-level θ's and the resulting higher-order factors (i.e., ξ and growth parameters) would not be reliably estimated.

The paper serves as two purposes. First, no prior paper has explicitly documented and reviewed the three popular longitudinal IRT models as well as their identifiability constraints with and without known item parameters. Including this information has profound didactic value for practitioners who wish to apply the models to their own data. Sample M*plus* code is provided in the Appendix for each model for readers' reference. Second, this paper is the first attempt to thoroughly compare and demonstrate the applicability of each of the discussed models. Even though these models can adequately capture changes in typical longitudinal measures, they are by no means exhaustive. A handful of other longitudinal models exist, such as the two-tier model (Cai, 2010), in which nuisance factors are introduced to account for residual dependencies between common items over time, or the item-level-growth-curve model (Paek, et al., 2016), in which growth rates for different items can differ and therefore be described and examined.

Regardless of chosen model, constructing and estimating growth using longitudinal IRT can improve the measurement of educational outcomes and, thus, provide educators with tools they need to better help students learn. Currently available software packages can estimate growth across a wide variety of measurement models (e.g., 1PL, 2PL, 3PL, unidimensional, multidimensional, and higher-order) and latent growth curve models (i.e., Equations (1) and (4)). Interested practitioners should be cognizant of the different estimation methods offered in each of the programs and to choose the method appropriate for the problem at hand, especially given complex models with many estimable parameters. For instance, the discussed analytic dimension reduction technique is only relevant to MML estimation approaches but not to the Bayesian MCMC estimation approach commonly used to estimate parameters of complex models. Software packages such as M*plus* may not automatically use a given dimension reduction unless the command file (or source script) is written with dimension reduction in mind. [5]. Hence, understanding the logic of dimension reduction can help with constructing the command file or

[5] Please see an example for the HO-IRT model at www.placeholder.com

script processed by the algorithm and greatly reduce computation time.

Although this didactic offers sufficient technical details for three popular longitudinal IRT models for researchers and practitioners to use those models in their own research, two relevant topics were outside the scope of the current discussion. First, LCG models with intrinsically nonlinear growth patterns were not discussed because this family of models is not currently included in a majority of software packages for LCG model estimation. An example of this kind of model is a "piece-wise growth curve model with unknown knots" (e.g., Kohli et al., 2015). Second, we have not discussed how to evaluate global model fit. Although most SEM software packages will output one or multiple absolute fit indices, few studies have examined appropriate cutoffs for these indices in determining adequate fit. Moreover, the DIC that is often used with MCMC can take different forms. The first-level conditional DIC provided by WinBUGS may not always provide the best estimates of model fit, whereas a second-level joint DIC might be more appropriate for multilevel IRT models (Zhang, Tao, & Wang, 2019). A thorough examination of model fit for longitudinal IRT models is needed to ensure credible conclusions drawn from any model-based results.

References

Andersen, E. B. (1985) Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of multivariate analysis*, 95(1), 1-22.

Baldwin, S. G., Baldwin, P., & Nering, M. L. (2007). *A comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9-25.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2018). lme4: Linear mixed-effects models using ‘Eigen’ and S4. R package version 1.1-20. http://CRAN.R-project.org/package=lme4

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471-492.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* Hoboken, NJ: Wiley-Interscience.

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581-612.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221-248.

Chen, F., Sousa, K., & West, S. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471-492.

Cronbach, L. J., & Furby, L. (1970). How we should measure “change”–or should we? *Psychological Bulletin*, *74*, 68–80.

Curtis, S. M. (2010). BUGS code for Item Response Theory. *Journal of Statistical Software*. Retrieved from https://www.jstatsoft.org/article/view/v036c01

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267–285.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620-639.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620–639.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (Eds.). (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*, 275-299.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423-436.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *159*, 505-513.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, *78*(1), 45-51.

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling*, *20*, 504-517.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Muller (Eds.), *Structural equation modeling: A second course* (pp. 171–196). Greenwood, CT: Information Age Publishing, Inc.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Hsieh, C.-A., von Eye, A. A., & Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the national youth survey. *Multivariate Behavioral Research*, *45*, 508–552.

Huang, H. Y. (2013). *Measuring Latent Growth under the Multilevel Higher-Order Item Response Theory Model.* Paper presented at 2013 Annual Meeting of National Council on Measurement in Education, San Francisco, CA.

Huang, H. Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement*, *39*, 362-372.

Huang, H. Y. & Wang, W. C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement*, *74*, 495-515.

Isiordia, M.,& Ferrer, E. (2018). Curve of factors model: A latent growth modeling approach for education research. *Educational and Psychological Measurement*,*78*, 203-231.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Eds.), *Handbook of Structural Equation Modeling*, Guilford Press.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136-153.

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie/Journal of Psychology*, *216*, 49–58.

Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. L. (2015). Fitting a linear-linear piecewise growth mixture model with unknown knots: A comparison of two common approaches to inference. *Psychological Methods*, *20*(2), 259-275.

Laird, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, *65*, 581-590.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2016). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, Advance online publication. http://dx.doi.org/10.1037/met0000075

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The handbook of multivariate experimental psychology, volume 2* (pp. 561–614). New York, NY: Plenum Press.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In Cudeck, R., du Toit, S., Sorbom, D., *Structural equation modeling: Present and future* (pp. 342-380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 137-175). Washington, DC: American Psychological Association.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Measurement*, *14*, 126–149.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* New York, NY: Chapman & Hall.

Meng, X.-L., & van Dyk, D. (1997). The EM algorithm–An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*, 511–567.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107-122.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide. Sixth edition.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (2010). *Bayesian analysis in Mplus: A brief introduction.* Available from http://www.statmodel.com/ download/introbayesversion%203.pdf.

Nydick, S. W. (2014). The sequential probability ratio and binary item response models. *Journal of Equational and Behavioral Statistics*, *39*, 202–230.

Paek, I., Li, Z., & Park., H. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate Behavioral Research*, *51*, 569-581.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proccedings of the 3rd Internaitonal Workshop on Distributed Statistical Computing*. March 20-22, Vienna, Austria.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). Retrieved from https://www.jstatsoft.org/article/view/v048i02

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*, 167-182.

Rodriguez, G. & Goldman, N. (1995). An assessment of estimation procedures for multi- level models with binary responses. *Journal of the Royal Statistical Society (Series A)*, *158*, 73-90.

Sheng, Y., & Wikle, C. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413-430.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society (Series B)*, *64*, 583-639.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*, 5-30.

Takane, Y, & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health outcomes. *Medical Care*, *44*, S39-S49.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*, 225-255.

U. S. Department of Education (2009). *Race to the top program executive summary.* Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.

Wang, C. (2014). Improving measurement precision of hierarchical latent traits using adaptive testing. *Journal of Equational and Behavioral Statistics*, *39*, 452–477.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 455-465.

Wang, C., & Nydick, S. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement*, *39*, 119–134.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research*, 281-324.

Wilson, M., Zheng, X., & McGuire, L. W. (2012). Formulating latent growth using an explanatory item response model appraoch. *Journal of Applied Measurement*, *13*, 1-22.

Sheng, Y., & Wikle, C. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413-430.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications. *Psychometrika*, *77*, 495-523.

Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement*, *51*, 18-38.

Zhang, X., Tao, J., & Wang, C. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement*.

Appendix: A Simulation Study

-

-

Appendix: A Simulation Study

Simulation Study

With the primary focus of comparing the three longitudinal IRT models, the current simulation is designed to assess the precision gain and/or loss in measuring individual growth when fitting data generated from the L-HO-IRT model with either the longitudinal MIRT or the longitudinal UIRT models. The L-HO-IRT model is used as the true model because it combines the key features of both L-MIRT and L-UIRT and, hence, it provides information that both L-MIRT and L-UIRT purport to estimate. For all simulation conditions, we estimated parameters with an MCMC algorithm. The number of time points, T, was fixed to be 4 across all conditions. We kept the number of time points fixed at $T = 4$ to be consistent with the operational data presented in the next section. Other constant factors across all conditions included (1) the simulee sample size fixed at 1,000 (Baker, 1998; de la Torre & Hong, 2010); (2) the number of dimensions, K, chosen to be 4; and (3) the test length fixed at $J = 60$ (so that $J_k = J/K = 60/4 = 15$ items measured each dimension).

Conditions

We included two manipulated factors in the simulation study. The residual variance of the growth model, σ_δ^2, was specified to be either 0.25 or 0.05. These two residual variances correspond to an intraclass correlation of 0.67 (for $\sigma_\delta^2 = 0.25$) or 0.91 (for $\sigma_\delta^2 = 0.05$). One would expect that a larger residual variance in estimation of the dependent variable (in this case, overall ability or ξ) would adversely affect the estimation precision of model parameters (Kohli, Hughes, Wang, Zopluoglu, & Davison, 2015). We also varied the bivariate correlation among the K lower-level factors to be either 0.3, 0.5, or 0.7. These correlations are consistent with those typically observed in real data. For instance, Yao (2014) fitted a four-dimensional simple-structure confirmatory item factor model to ASVAB (Armed Services Vocational Aptitude Battery) data and found the correlations among four factors varied between 0.3 and 0.7. A higher correlation among factors should lead to better recovery of overall ability and its growth trajectories. For each of the 6 conditions, we conducted $R = 100$ replications and aggregated results (using the median rather than the mean) across the replications.

Item Parameter Generation

Across all simulations, item parameters were generated according to the IRT model described in Equation (3). Because M*plus* parameterizes the IRT model with thresholds rather than difficulties, the estimates from M*plus* were a_{J_k} and d_{J_k}, where $d_{J_k} = a_{J_k} b_{J_k}$. Moreover, due to simple structure, and because J_k items loaded onto each dimension, the total number of items would always be $J = J_k \times K$. Different item parameter sets were generated for every replication within a condition and then estimated with each of the three models. All item slope parameters (a_{J_k}) were generated to be independent and identically distributed according to the log-normal distribution with a log-mean of 0 and log-standard deviation of 0.4 (which is approximately a mean of $\mu_a = 1.083$ and a standard deviation of $\sigma_a = 0.451$). All item difficulty parameters (b_{J_k}) were generated to be independent and identically distributed according to the normal distribution with a mean of $\mu_b = 0.2$ and a standard deviation of $\sigma_b = 1.3$).

Person Parameter Generation

Person parameters were generated as follows. We first simulated a set of linear model parameters, π_0 and π_1, and then used the linear parameters to generate higher-order abilities. The intercept parameter, π_0, was generated to be normally distributed with mean of $\mu_{\pi_0} = 0.0$ and variance of $\sigma^2_{\pi_0} = 0.5$. The slope parameter, π_1, was generated to be normally distributed with mean $\mu_{\pi_1} = 0.25$ and variance $\sigma^2_{\pi_1} = 0.01$. Given π_{0i} and π_{1i} for person i, higher-order person parameters were then set to

$$\xi_{it} = \pi_{0i} + \pi_{1i} \times (t - 1) + \delta_{it},$$

where $\delta_{it} \sim N(\mu_\delta = 0.0, \sigma^2_\delta)$ with $\sigma^2_\delta = 0.05$ or 0.25. Next, let $\theta^{(t)}_{ik}$ be the i^{th} persons lower-order ability parameter for dimension k at time point t. Then $\theta^{(t)}_{ik} = \lambda \xi^{(t)}_i + \epsilon^{(t)}_{ik}$, where $\epsilon^{(t)}_{ik} \sim N(\mu_\epsilon = 0.0, \sigma^2_\epsilon = 1 - \lambda^2)$. This parameterizations is the same as that described in de la Torre and Hong (2009). The value of λ were selected to be either $\sqrt{0.3}$, $\sqrt{0.5}$, or $\sqrt{0.7}$. Notice that in this data generation method, the relationship between the lower-order and higher-order person parameters were identical across dimensions and not assumed to change over time, simply to keep the generation scheme clear without loss of generality. In contrast, when fitting the model, the estimated loadings were allowed to vary across dimensions, but the relationship between the higher-order factor and a particular lower-order factor was assumed to not change over time, ensuring longitudinal invariance.

Model Fitting

For each simulation, we generated a response matrix of size $N \times (JT)$, where each simulee responded to each item across all time points. The Markov chain length was 10,000 with the first 5,000 iterations treated as burn-in. The chain length was selected such that the Proportion Scale Reduction (PSR) factor of all model parameters would be close to 1 (and all were well below 1.05).

As described earlier, we fit three models to each generated response. To avoid the need for post-hoc scale transformation, the model constraints were chosen carefully to closely match the true values. Specifically, when fitting the L-HO-IRT model, the first-level residual variances $\sigma^2_{\epsilon^1_{ik}} = c^1_k$ were fixed at $1 - \lambda^2_k$ (which is consistent with our data generation method). When fitting the L-MIRT model, the variances of the residuals across all dimensions at time point 1, $\sigma^2_{\delta^{(1)}_{ik}}$, were fixed at $\lambda^2_k \sigma^2_{\delta^{(t)}_i} + (1 - \lambda^2_k)$ (which is essentially the variance of $\upsilon^{(t)}_{ik}$ in Equation **??**). When fitting the L-UIRT model, the variance of the residuals at time point 1 was fixed at $\sigma^2_{\delta^{(1)}_i}$. In each case, $\sigma^2_{\delta^{(1)}_i}$ always took the generating value of either 0.05 or 0.25 as described in the previous section[1].

Evaluation Criteria

The evaluation criteria for parameter recovery include the average bias, the Mean Squared Error (MSE), and the correlation between generated and estimated parameters. We also recorded the number of converged replications, and only successful replications were included in the final analysis and all of the tables and graphs in the following sections[2].

The reported person parameters depended on the model. When fitting the L-HO-IRT

[1] This choice is made for the sake of evaluating parameter recovery. In practice, the value of $\sigma^2_{\delta^{(1)}_i}$ can be set at any constant.

[2] Technically, we calculated the MSE using the equation $\frac{\sum(\hat{\gamma}-\gamma)^2}{N}$, where γ is the desired parameter, which is estimated by $\hat{\gamma}$, and the summation was taken across the entire sample of N simulees. These statistics were then aggregated, using the median rather than the mean, across the set of 100 replications and then averaged across the desired conditions.

model, we reported the estimation precision of both the ξ's and the θ's. With respect to the L-MIRT model, we did not estimate any of the higher-order parameters (the ξ's), and, thus, only reported the estimation precision of the θ's. Finally, when fitting the L-UIRT model, we only estimated a single unidimensional trait, and, thus, only reported the estimation precision of the ξ's (which represents the best guess of that trait). The reported item parameters, in contrast, were identical for all three models: the recovery of the a's and b's parameters were reported and compared across all of the models.

Finally, we reported the parameter estimates and factor scores as related to the growth part of the model. In particular, for to the L-HO-IRT model, we reported the estimation precision of individual intercept and slope parameters with respect to both the higher-order factor, π_{0i} and π_{1i}, and the domain factors, $\lambda_k \pi_{0i}$ and $\lambda_k \pi_{1i}$. For the longitudinal MIRT model, we only reported the precision of the individual intercept and slope parameters with respect to the domain factors (as the higher-order growth part of the model is never estimated), whereas for the longitudinal UIRT model, we only reported the precision of individual intercept and slope parameters with respect to the higher-order, general abilities. Table 1 summarizes the parameter of interest for different models. Note that for the L-MIRT model, both $\hat{\text{cov}}(\boldsymbol{\pi}_{0i})$ and $\hat{\text{cov}}(\boldsymbol{\pi}_{1i})$ denote the K-by-K covariance matrices of individual intercepts and individual slopes respectively. Although the full covariance matrix is estimated, only the diagonal elements need to be compared to the "true" variances of the individual intercept and slope parameters at the domain level (i.e.,$\lambda_k^2 \sigma_{\pi_{0i}}^2$ and $\lambda_k^2 \sigma_{\pi_{1i}}^2$).

Table 1
Target parameters of interest for different models

	True parameters	Fitted model		
		L-HO-IRT	L-MIRT	L-UIRT
Person	ξ_i, θ_{ik}	$\hat{\xi}_i$, $\hat{\theta}_{ik}$	$\hat{\theta}_{ik}$	$\hat{\xi}_i$
Individual growth	π_{0i}, π_{1i} for ξ $\lambda_k \pi_{0i}$, $\lambda_k \pi_{1i}$ for θ_k	$\hat{\pi}_{0i}$, $\hat{\pi}_{1i}$ $\hat{\lambda}_k \hat{\pi}_{0i}$, $\hat{\lambda}_k \hat{\pi}_{1i}$	$\hat{\pi}_{0i}^k$, $\hat{\pi}_{1i}^k$	$\hat{\pi}_{0i}$, $\hat{\pi}_{1i}$
Population growth	β_1 $\lambda_k \beta_1$ $\sigma_{\pi_{0i}}^2$, $\sigma_{\pi_{1i}}^2$, $\sigma_{\delta_i^{(t)}}^2$	$\hat{\beta}_1$ $\hat{\lambda}_k \hat{\beta}_1$ $\hat{\sigma}_{\pi_{0i}}^2$, $\hat{\sigma}_{\pi_{1i}}^2$, $\hat{\sigma}_{\delta_i^{(t)}}^2$	$\hat{\beta}_1^k$ $\hat{\text{cov}}(\boldsymbol{\pi}_{0i})$, $\hat{\text{cov}}(\boldsymbol{\pi}_{1i})$, $\hat{\sigma}_{\delta_{ik}^{(t)}}^2$	$\hat{\beta}_1$ $\hat{\sigma}_{\pi_{0i}}^2$, $\hat{\sigma}_{\pi_{1i}}^2$, $\hat{\sigma}_{\delta_i^{(t)}}^2$
Item	a, b	$\hat{a}$, $\hat{b}$	$\hat{a}$, $\hat{b}$	$\hat{a}$, $\hat{b}$

Simulation Results

Table 2 summarizes the number of failed (in terms of model convergence) replications (out of 100) per condition per model. As seen in Table 2, the convergence rate is above 95% across all cells. The failed replications are mainly due to non-mixed responses for certain items. Using only the converged conditions, we describe parameter recovery in three subsections. We first explain how overall and domain abilities were recovered when using each of the three models. We then explain how the item parameters were recovered. Finally, we explain how the individual growth parameters were recovered.

Table 2

Number of failed replications per condition per model out of 100 total replications

σ_δ^2	r	Longitudinal HO-IRT	Longitudinal MIRT	Longitudinal UIRT
0.05	0.3	2	2	2
	0.5	2	2	2
	0.7	4	3	3
0.25	0.3	0	0	0
	0.5	5	5	5
	0.7	3	4	3

Overall and Domain Abilities Recovery

Figure 1 displays the bias (top panels), MSE (middle panels), and correlation (bottom panels) for the three correlation conditions (left panels) and two residual level conditions (right panels) when comparing L-HO-IRT and L-UIRT estimation recovery on the higher-order factor, ξ. The purple symbols present the results from the higher-order model, whereas the green symbols present the results from the unidimensional model. Points within a plot indicate the average bias, correlation, and MSE for a given person parameter at each of the four time points. Average bias was calculated by subtracting the true/generated parameters from the estimated parameters, averaging across all persons, and then taking the median across replications.

When looking at the results from the longitudinal HO-IRT model alone, one can find two, fairly obvious, trends as pertaining to the conditions presented in Figure 1. First and unsurprisingly, increasing the correlations among the lower-order factors leads to better recovery of the ξ's, as reflected by the lower average MSE and higher average correlation when comparing true versus estimated ξ. Second and also consistent with our expectation, having smaller residual variances yields better recovery of ξ. With respect to the L-HO-IRT model, the bias of the ξ estimates were fairly close to 0 across all conditions.

Unlike results from the longitudinal HO-IRT model, fitting the longitudinal UIRT model to HO-IRT data led to worse estimation precision. Unsurprisingly, as shown in the left panels of Figure 1, an increase in the model-specified correlation between the lower-order traits led to an concurrent decrease in the bias and MSE of the estimated abilities as well as an increase in the correlation between estimated and true abilities. This result is to be expected because a higher correlation between lower-order traits results in a stronger higher-order factor and a test that is essentially unidimensional. However, contrary to the pattern observed from estimating ξ from the L-HO-IRT model, the absolute value of bias and MSE actually tended to increase at later time points when estimating abilities with the L-UIRT model. This outcome is due to errors contained in the slope parameter estimates propogating through to all ξ estimates when $t \geq 2$. The correlation, however, does not seem to be as sensitive as the bias and MSE as an evaluation criterion. Unlike the bias and MSE, (1) correlations do not capture any changes or distortion in the scale of the latent trait, and (2) larger variability in true parameters might yield deceptively high correlations for larger values of t.

The trend in the left panels (when comparing different true correlations among the domain abilities) also appear in the right panels (when comparing different specified residual variances). Interestingly, the L-HO-IRT and L-UIRT models only start to diverge in their estimating precision of ξ when $t \geq 2$. Therefore, the differential precision of the different estimating models appear to be mainly due to the recovery of the slope parameters. In fact, when the true residual variance of ξ is large, the bias of the estimated person-specific slope parameters is smaller (see Tables 4 and 6 for additional details). Ultimately, when the true data is generated from the longitudinal HO-IRT model, mistakenly fitting the simpler longitudinal UIRT appears to result in a precision loss in recovering overall abilities, and this loss becomes worse over time. Also this precision loss is more salient when the correlations among the domain-level abilities are low or when the residual variances in the true ξ's are small. A scatterplot of the true versus estimated

ξ's (which is not reported here to save space) reveals that fitting the L-UIRT model to HO-IRT data yielded aberrantly low variability in the estimated ξ's.

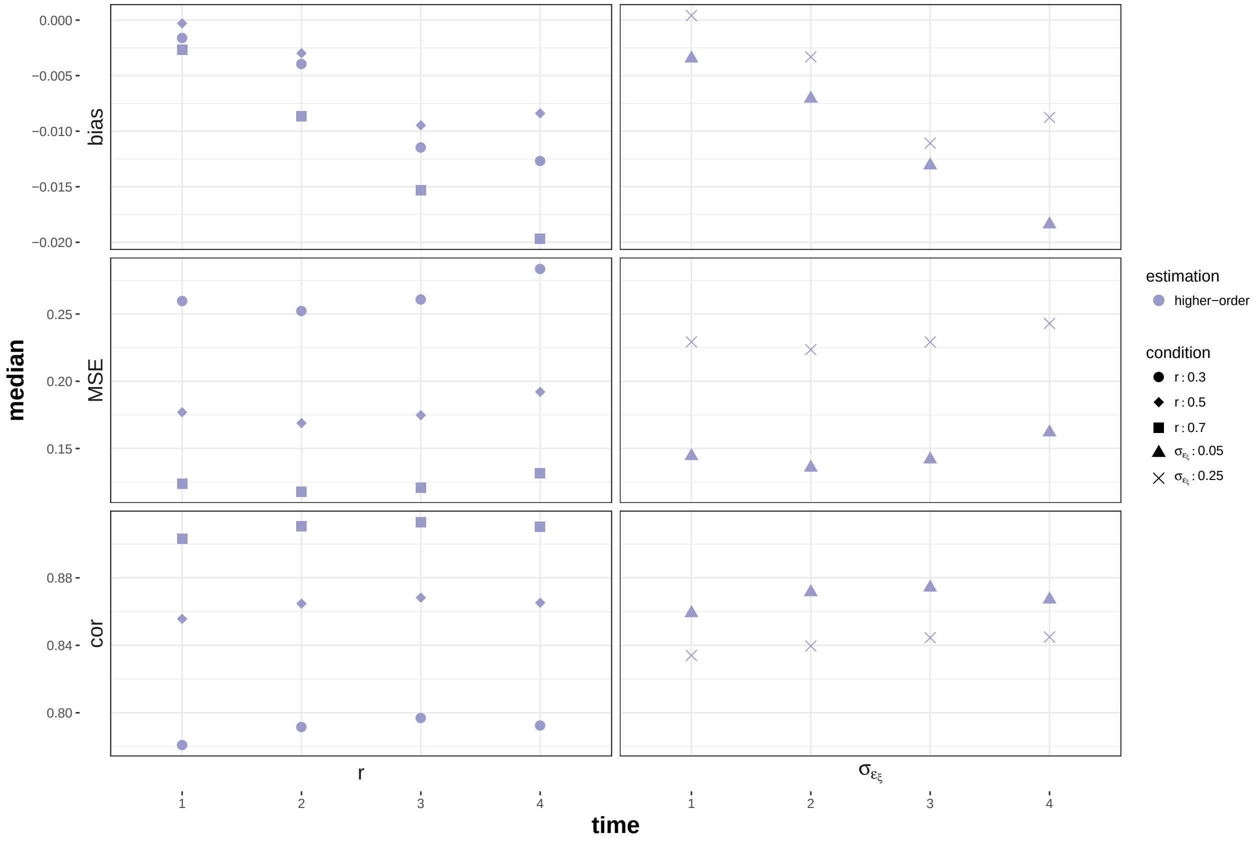

Figure 1. The bias, MSE, and correlation when comparing true and estimated ξ across time from both longitudinal HO-IRT and longitudinal UIRT models. The results were aggregated within either correlation (left three panels) or residual variance (right three panels) conditions. Note that the top plot displays bias results when subtracting the true value of a person parameter from its estimated value.

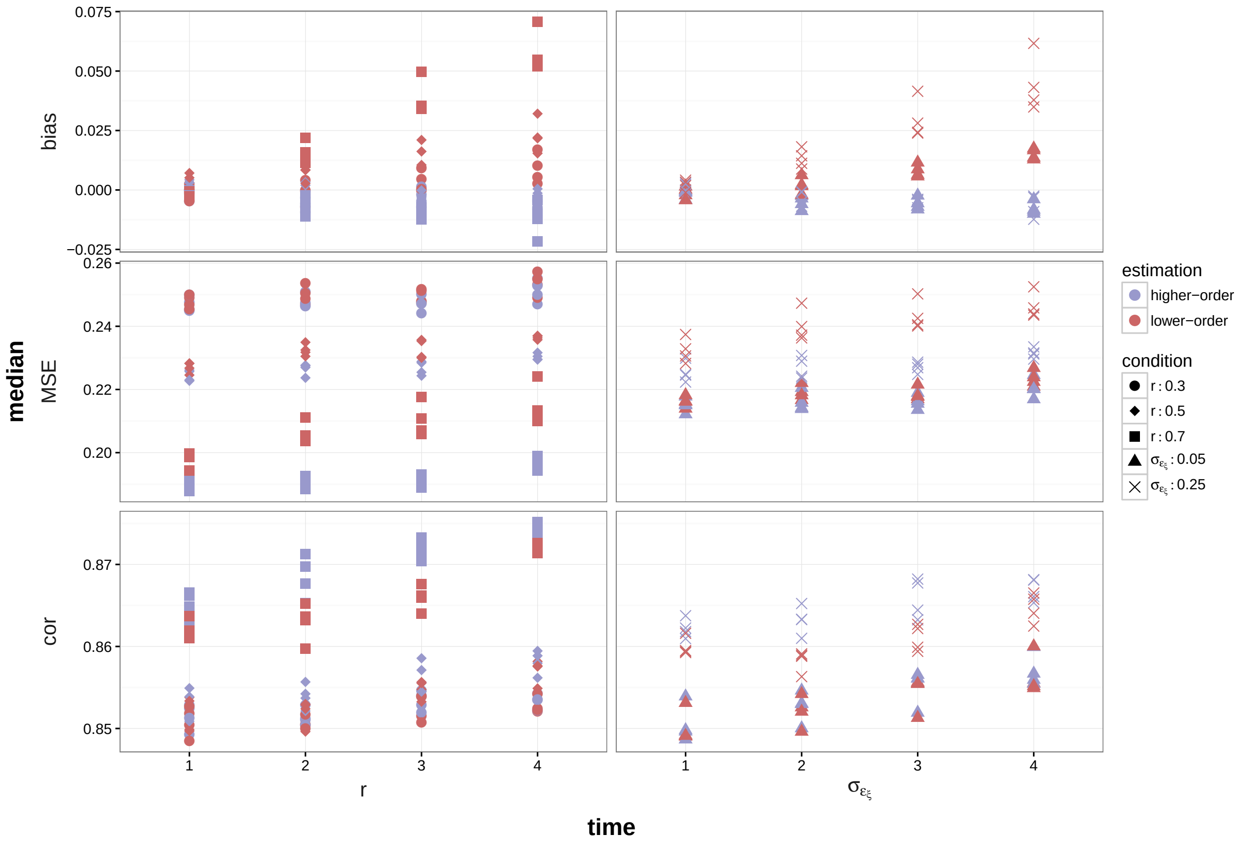

Figure 2. The bias, MSE, and correlation when comparing true and estimated θ across time from both longitudinal HO-IRT and longitudinal MIRT models. The results were aggregated within either correlation (left three panels) or residual variance (right three panels) conditions. Note that the top plot displays bias results when subtracting the true value of a person parameter from its estimated value.

Figure 2 presents the recovery of θ's from both the longitudinal HO-IRT model and the longitudinal MIRT model, aggregated within either correlation condition (left panels) or residual variance condition (right panels). As shown in Figure 2, the person-specific θ's were actually recovered pretty well when using either model under nearly all conditions, as demonstrated in the small y-axis range in the plot. Moreover, points of the same shape but different colors are always fairly close in Figure 2, also indicating that both models recover lower-order abilities with similar precision. However, one can still spot a few trends from Figure 2, indicating some differential performance in the two models. First, an increase in the correlation among lower-order factors yielded a smaller MSE and higher correlation, but at the cost of a slight increase in bias when $t \geq 2$, which is commonly known as the bias variance tradeoff. Second, a larger residual variance in the higher-order parameters appears to have adversely affected estimation recovery when using the L-MIRT model, as shown by the increased bias and MSE of θ's over time for the red X's on the right side of the plot. In contrast to the estimation precision when using L-MIRT, the effect of a large residual variance on the estimation precision of the L-HO-IRT model for $t \geq 2$ was almost negligible. One explanation for the lack of a time-dependent effect of residual variance on the estimation precision of the L-HO-IRT model is that the residual variance was specified to be on the higher-order trait, so that the effect of a larger residual variance of ξ on the estimation of the lower order abilities was indirect, and the ultimate cost on θ recovery was small.

Item Parameters Recovery

Table 3 displays the bias of the estimated item parameters within each condition when estimating parameters using each of the three models. The most striking observation from Table 3 is that the $\hat{a}$-parameters show extreme positive bias when fitting the longitudinal UIRT model in the condition when the residual variance is small and the correlations among lower-order factors are small. In this condition, the set of items displays extreme multidimensionality, due to the small correlation among the domain-level abilities. Moreover, a test with small residual variance and small correlations among lower-order factors would be very close to the ideal L-HO-IRT generating model. Therefore, the positive bias in the a-parameter estimates from the L-UIRT model is probably due to the very low person variability relative to item variability when $\sigma_\delta^2 = .05$. $\hat{a}$-parameter estimates became negatively biased, but with much smaller magnitude, when the residual variance was large. This observation is consistent with the pattern observed in Figure 1, in that ξ is recovered poorly when both σ_δ^2 and r are low.

When examining the results from the longitudinal MIRT model, one can notice that the bias in the item parameter estimates are generally acceptable even though they are still larger, in magnitude, than results from the L-HO-IRT model. The L-MIRT model contains more parameters, and hence, in theory, it should also adequately fit the response matrix. However, the additional, unnecessary parameters in L-MIRT results in increased difficulty in estimating item properties, which yields more visible bias. Even so, the bias in estimating the item parameters with the L-MIRT model only becomes unacceptable when $\sigma_\delta^2 = 0.25$ and $r = 0.7$. This particular result would be expected, as high correlations among the dimensions has been known to lead to estimation difficulty of the MIRT model item parameters (e.g., Chen & Wang, 2016; Wang & Nydick, 2015). Moreover, introducing noise, via high residual variances, would further exacerbate the difficulty the L-MIRT model has in estimating item parameters. In contrast to the L-UIRT and L-MIRT models, the item parameter estimates from the L-HO-IRT model are all nearly unbiased, indicating that the MCMC algorithm adequately recovered the true item parameters when the model was correctly specified.

Table 4 displays the corresponding MSE of the item parameters estimates for those conditions outlined earlier. Compared to Table 3, the same patterns persist in nearly all of the cells in Table 4, suggesting that bias contributes the most to the actual MSE values in all cells of the table. Moreover, for the L-HO-IRT model, the bias and MSE value are so small that the differences across the conditions up to two (or three in some cases) decimal places are negligible.

Table 3

Average bias for estimates of item parameters for all manipulated conditions.

		Longitudinal HO-IRT					Longitudinal MIRT					Longitudinal UIRT				
σ^2_δ	r	d	a_1	a_2	a_3	a_4	d	a_1	a_2	a_3	a_4	d	a_1	a_2	a_3	a_4
0.05	0.3	0.007	-0.023	-0.010	-0.001	-0.019	0.005	-0.034	-0.043	-0.030	-0.048	0.030	0.805	0.829	0.691	0.726
	0.5	0.000	-0.020	-0.010	0.000	-0.012	0.000	-0.035	-0.057	-0.036	-0.055	0.010	0.689	0.676	0.709	0.676
	0.7	0.009	-0.005	-0.004	0.009	-0.008	0.009	-0.053	-0.085	-0.064	-0.082	0.019	0.553	0.553	0.519	0.556
0.25	0.3	0.007	-0.011	0.000	-0.003	-0.022	0.006	-0.038	-0.056	-0.043	-0.059	0.026	-0.110	-0.130	-0.106	-0.101
	0.5	0.007	-0.016	-0.003	0.002	-0.016	0.007	-0.076	-0.083	-0.082	-0.092	0.012	-0.044	-0.068	-0.062	-0.058
	0.7	0.003	-0.029	-0.010	0.044	0.005	0.001	-0.168	-0.166	-0.135	-0.152	0.016	-0.012	-0.010	-0.018	-0.004

Table 4

MSE for estimates of item parameters for all manipulated conditions.

		Longitudinal HO-IRT					Longitudinal MIRT					Longitudinal UIRT				
σ^2_δ	r	d	a_1	a_2	a_3	a_4	d	a_1	a_2	a_3	a_4	d	a_1	a_2	a_3	a_4
0.05	0.3	0.014	0.010	0.009	0.010	0.013	0.014	0.010	0.010	0.011	0.014	0.078	0.712	0.742	0.526	0.557
	0.5	0.013	0.010	0.010	0.010	0.010	0.014	0.011	0.014	0.012	0.014	0.056	0.513	0.484	0.539	0.496
	0.7	0.015	0.012	0.012	0.014	0.014	0.015	0.016	0.017	0.016	0.019	0.038	0.338	0.341	0.295	0.346
0.25	0.3	0.013	0.008	0.011	0.009	0.010	0.013	0.010	0.011	0.011	0.013	0.072	0.049	0.054	0.036	0.041
	0.5	0.012	0.009	0.010	0.009	0.011	0.015	0.014	0.019	0.017	0.024	0.050	0.020	0.023	0.021	0.031
	0.7	0.014	0.015	0.013	0.012	0.014	0.015	0.043	0.040	0.031	0.042	0.043	0.014	0.018	0.016	0.017

Growth Parameters Recovery

Table 5

Average bias of the individual intercept and slope from the three models under all manipulated conditions

Conditions		Growth of ξ		Growth of $\boldsymbol{\theta}$							
Longitudinal HO-IRT											
σ^2_δ	r	$\hat{\pi}_{0i}$	$\hat{\pi}_{1i}$	$\hat{\lambda}_1\hat{\pi}_{0i}$	$\hat{\lambda}_1\hat{\pi}_{1i}$	$\hat{\lambda}_2\hat{\pi}_{0i}$	$\hat{\lambda}_2\hat{\pi}_{1i}$	$\hat{\lambda}_3\hat{\pi}_{0i}$	$\hat{\lambda}_3\hat{\pi}_{1i}$	$\hat{\lambda}_4\hat{\pi}_{0i}$	$\hat{\lambda}_4\hat{\pi}_{1i}$
0.05	0.3	-0.003	-0.004	-0.001	-0.002	-0.001	-0.005	-0.001	-0.006	-0.001	-0.004
	0.5	-0.002	-0.004	-0.001	-0.003	-0.001	-0.006	-0.001	-0.005	-0.001	-0.005
	0.7	-0.005	-0.007	-0.004	-0.006	-0.004	-0.006	-0.004	-0.010	-0.004	-0.006
0.25	0.3	0.003	-0.003	0.001	-0.002	0.001	-0.008	0.001	-0.004	0.001	-0.002
	0.5	0.004	-0.004	0.003	-0.003	0.003	-0.004	0.003	-0.008	0.003	-0.006
	0.7	0.006	-0.006	0.005	-0.005	0.005	-0.013	0.005	-0.019	0.005	-0.008
		L-UIRT		Longitudinal MIRT							
σ^2_δ	r	$\hat{\pi}_{0i}$	$\hat{\pi}_{1i}$	$\hat{\pi}^1_{0i}$	$\hat{\pi}^1_{1i}$	$\hat{\pi}^2_{0i}$	$\hat{\pi}^2_{1i}$	$\hat{\pi}^3_{0i}$	$\hat{\pi}^3_{1i}$	$\hat{\pi}^4_{0i}$	$\hat{\pi}^4_{1i}$
0.05	0.3	-0.003	-0.183	-0.003	0.002	-0.003	0.002	-0.003	0.002	-0.002	0.000
	0.5	-0.004	-0.155	-0.003	0.006	-0.003	0.004	-0.003	0.004	-0.003	0.002
	0.7	-0.007	-0.122	-0.007	0.013	-0.008	0.011	-0.008	0.010	-0.007	0.007
0.25	0.3	0.001	-0.122	-0.001	0.005	-0.001	0.000	-0.001	0.007	-0.001	-0.001
	0.5	0.001	-0.085	-0.001	0.018	-0.001	0.011	-0.001	0.007	-0.001	0.005
	0.7	0.004	-0.055	0.002	0.041	0.002	0.026	0.002	0.023	0.002	0.025

Table 5 and Table 6 display the bias in the individual growth parameter estimates given each of the three models and across all of the conditions. In general, when estimating parameters using the true L-HO-IRT model, the growth parameters (that is the individual slope and intercept parameters) were all recovered well. The same general trend continues to hold as in the previous two sections, in that higher correlations among the domain abilities and lower residual variances in the higher-order trait leads to better estimation accuracy of the L-HO-IRT model parameters. In contrast to the L-HO-IRT model, the L-MIRT model and L-UIRT model only estimated growth trajectories at one level of the model. As shown in Tables 5 and 6, the L-MIRT model resulted in acceptable recovery of the growth trajectories on the lower-order θ's. However, the magnitude of the bias and MSE are higher, although only slightly higher, than estimates from the L-HO-IRT model. As described in the previous section, over-fitting the data with a less parsimonious model continues to result in valid inferences on the individual growth parameters.

On the other hand, the longitudinal UIRT model resulted in quite a bit of negative bias

Table 6

MSE of the individual intercept and slope from the three models under all manipulated conditions

Conditions		Growth of ξ		Growth of $\boldsymbol{\theta}$							
				Longitudinal HO-IRT							
σ^2_δ	r	$\hat{\pi}_{0i}$	$\hat{\pi}_{1i}$	$\hat{\lambda}_1\hat{\pi}_{0i}$	$\hat{\lambda}_1\hat{\pi}_{1i}$	$\hat{\lambda}_2\hat{\pi}_{0i}$	$\hat{\lambda}_2\hat{\pi}_{1i}$	$\hat{\lambda}_3\hat{\pi}_{0i}$	$\hat{\lambda}_3\hat{\pi}_{1i}$	$\hat{\lambda}_4\hat{\pi}_{0i}$	$\hat{\lambda}_4\hat{\pi}_{1i}$
0.05	0.3	0.171	0.010	0.051	0.003	0.051	0.003	0.051	0.003	0.051	0.003
	0.5	0.112	0.009	0.056	0.005	0.056	0.005	0.056	0.005	0.056	0.005
	0.7	0.080	0.009	0.056	0.006	0.056	0.006	0.056	0.006	0.056	0.006
0.25	0.3	0.194	0.010	0.058	0.003	0.058	0.003	0.058	0.003	0.058	0.003
	0.5	0.141	0.009	0.071	0.005	0.071	0.005	0.071	0.005	0.071	0.005
	0.7	0.113	0.009	0.079	0.007	0.079	0.007	0.079	0.007	0.079	0.007
		L-UIRT		Longitudinal MIRT							
σ^2_δ	r	$\hat{\pi}_{0i}$	$\hat{\pi}_{1i}$	$\hat{\pi}^1_{0i}$	$\hat{\pi}^1_{1i}$	$\hat{\pi}^2_{0i}$	$\hat{\pi}^2_{1i}$	$\hat{\pi}^3_{0i}$	$\hat{\pi}^3_{1i}$	$\hat{\pi}^4_{0i}$	$\hat{\pi}^4_{1i}$
0.05	0.3	0.353	0.043	0.053	0.004	0.053	0.004	0.052	0.004	0.053	0.003
	0.5	0.262	0.033	0.058	0.006	0.057	0.005	0.058	0.006	0.058	0.005
	0.7	0.178	0.024	0.059	0.007	0.058	0.007	0.058	0.007	0.059	0.007
0.25	0.3	0.272	0.024	0.060	0.005	0.060	0.005	0.060	0.005	0.061	0.005
	0.5	0.184	0.016	0.079	0.010	0.078	0.010	0.078	0.010	0.079	0.009
	0.7	0.131	0.012	0.114	0.023	0.106	0.019	0.106	0.019	0.109	0.020

when $\sigma^2_\delta = .05$ (which also explains the large bias of $\hat{\xi}$ in Figure 1). Furthermore, the MSE of the intercept and slope estimates both decreased as the correlation between the domain abilities, r, increased. These findings suggest that when a test exhibits multidimensionality, fitting a simpler unidimensional model will lead to biased conclusions regarding individual growth trajectories on the general ability.

Finally, Table 7 reports the median bias of the population growth parameter estimates from the three models. Unlike the previous tables, population growth is simply the population average and variance of the growth parameters across all possible simulees. Note that population growth is a parameter in the model and is not simply estimated by taking the average or variance of the sample estimates of the individual slopes and intercepts. With regard to both the L-HO-IRT and the L-MIRT models, all population growth parameters resulted in small bias with the exception of one cell. When $r = 0.7$ and $\sigma^2_\delta = 0.25$ (the highest correlation between the domain abilities and the largest residual variance), the L-MIRT model resulted in fairly large positive bias for the variance of the intercepts. In these conditions, the $\hat{\theta}$'s would tend to be more dispersed than their true values. This observation echoes results from Table 3.

Unsurprisingly, given previous results, the L-UIRT model leads to poorer recovery of the parameters, as reflected by the large negative bias of β_1 and $\sigma^2_{\pi_{0i}}$. Again, this observation is consistent with the previous results, and implies that the L-UIRT model tends to dramatically underestimate the actual variability of the latent trait.

Table 7

Median bias of the population growth parameters from the three models under all manipulated conditions

Conditions		Longitudinal HO-IRT			Longitudinal UIRT								
σ^2_δ	r	β_1	$\sigma^2_{\pi_{0i}}$	$\sigma^2_{\pi_{1i}}$	β_1	$\sigma^2_{\pi_{0i}}$	$\sigma^2_{\pi_{1i}}$						
0.05	0.3	-0.002	0.012	0.003	-0.184	-0.464	-0.009						
	0.5	0.001	-0.006	0.003	-0.155	-0.427	-0.009						
	0.7	-0.002	-0.010	0.000	-0.122	-0.370	-0.008						
0.25	0.3	-0.001	-0.002	0.007	-0.121	-0.370	-0.007						
	0.5	-0.003	-0.007	0.002	-0.086	-0.280	-0.006						
	0.7	0.000	0.021	0.002	-0.053	-0.193	-0.005						
						Longitudinal MIRT							
σ^2_δ	r	β^1_1	β^2_1	β^3_1	β^4_1	$\sigma^2_{\pi^1_{0i}}$	$\sigma^2_{\pi^1_{1i}}$	$\sigma^2_{\pi^2_{0i}}$	$\sigma^2_{\pi^2_{1i}}$	$\sigma^2_{\pi^3_{0i}}$	$\sigma^2_{\pi^3_{1i}}$	$\sigma^2_{\pi^4_{0i}}$	$\sigma^2_{\pi^4_{1i}}$
0.05	0.3	0.004	0.000	0.000	0.003	0.016	0.005	0.030	0.007	0.015	0.008	0.023	0.006
	0.5	0.005	0.005	0.004	0.002	0.018	0.007	0.023	0.008	0.017	0.008	0.029	0.007
	0.7	0.010	0.015	0.007	0.010	0.037	0.008	0.068	0.009	0.046	0.009	0.052	0.008
0.25	0.3	0.003	-0.001	0.007	-0.001	0.024	0.010	0.029	0.011	0.022	0.012	0.035	0.012
	0.5	0.013	0.010	0.008	0.006	0.046	0.018	0.060	0.021	0.054	0.020	0.057	0.018
	0.7	0.033	0.035	0.021	0.024	0.178	0.035	0.188	0.036	0.145	0.033	0.154	0.033